Spatial Maps of Prostate Cancer Transcriptomes Reveal an Unexplored Landscape of Heterogeneity

Emelie Berglund, Jonas Maaskola, Niklas Schultz, Maja Marklund, Stefanie Friedrich, Joseph Bergenstråhle, Firas Tarish, Anna Tanoglidi, Sanja Vickovic, Ludvig Larsson, Fredrik Salmén, Christoph Ogris, Karolina Wallenborg, Jens Lagergren, Patrik Ståhl, Erik Sonnhammer, Thomas Helleday and Joakim Lundeberg

Supplementary information

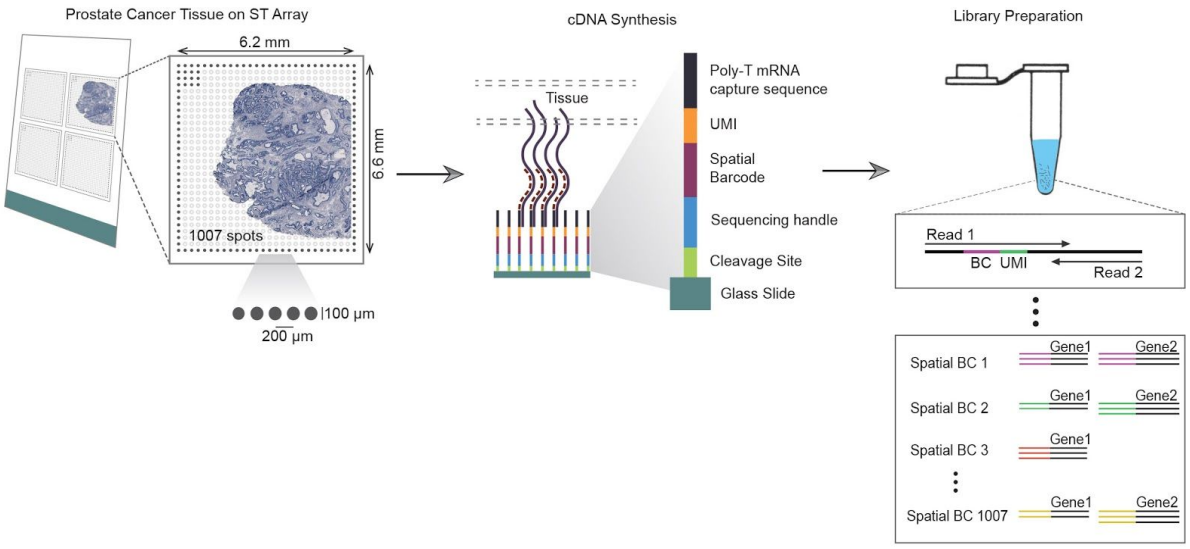
- Supplementary Figures

- Supplementary Table

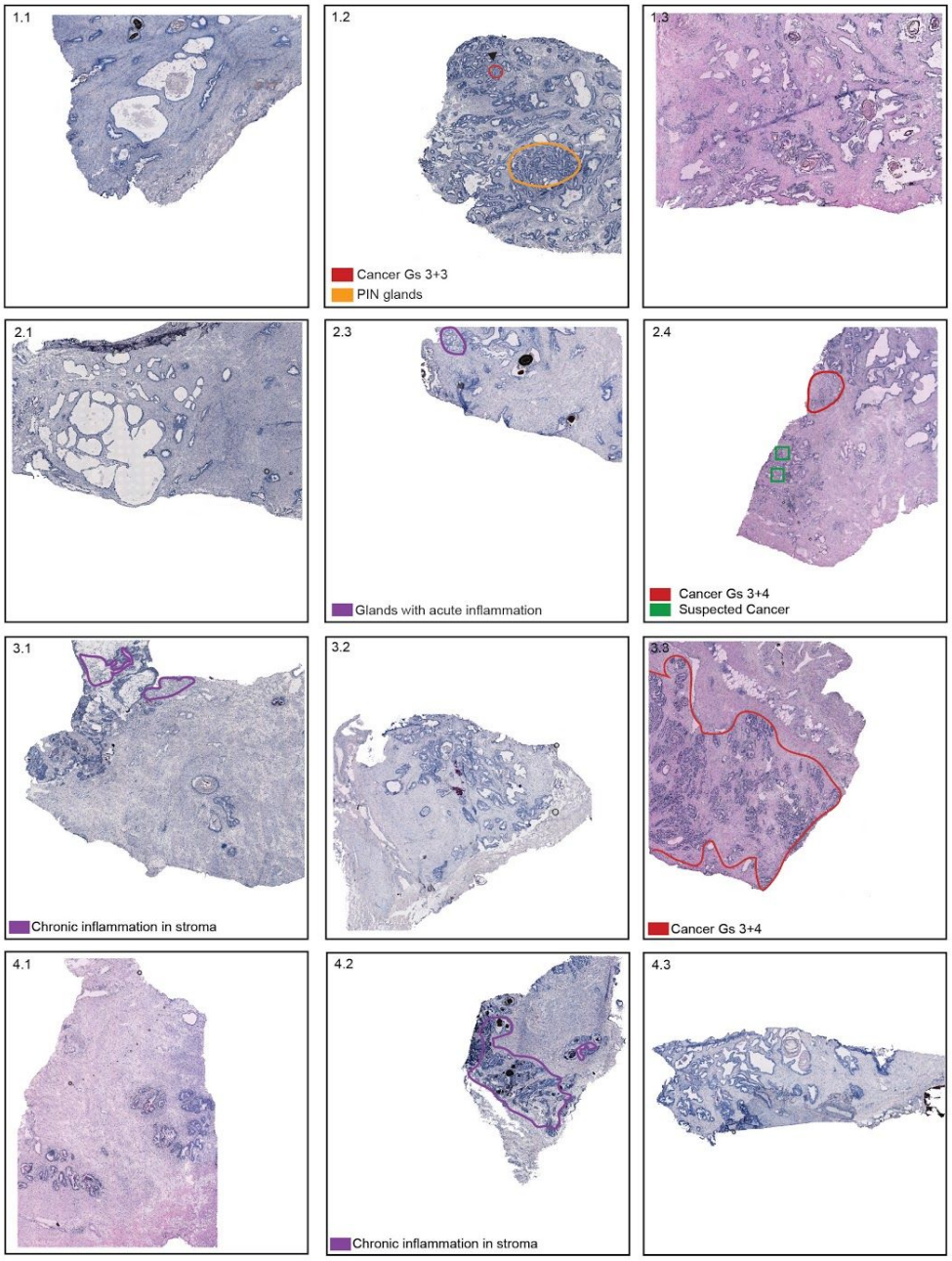
- Supplementary Methods

- Supplementary References

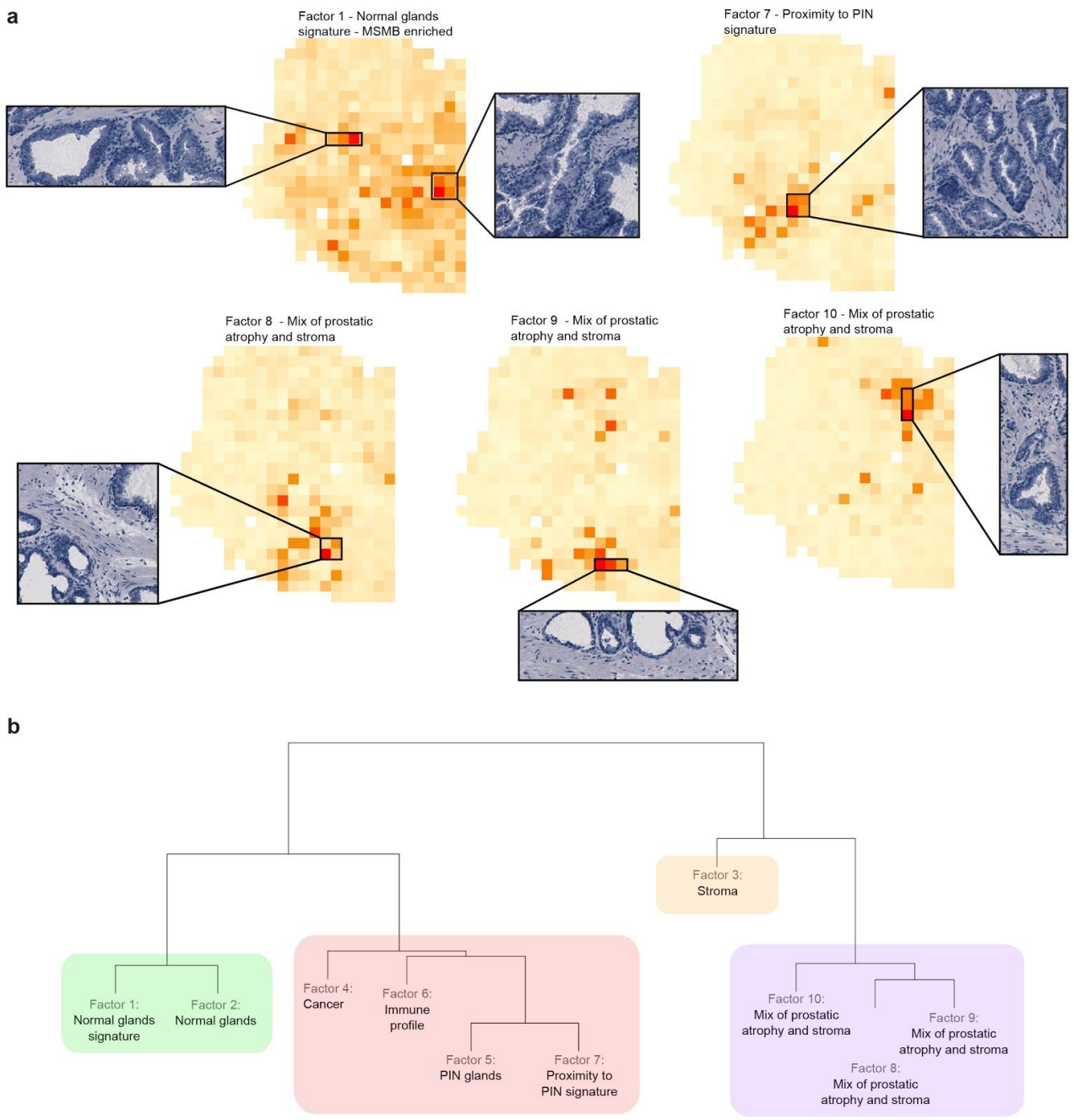
Supplementary Figures



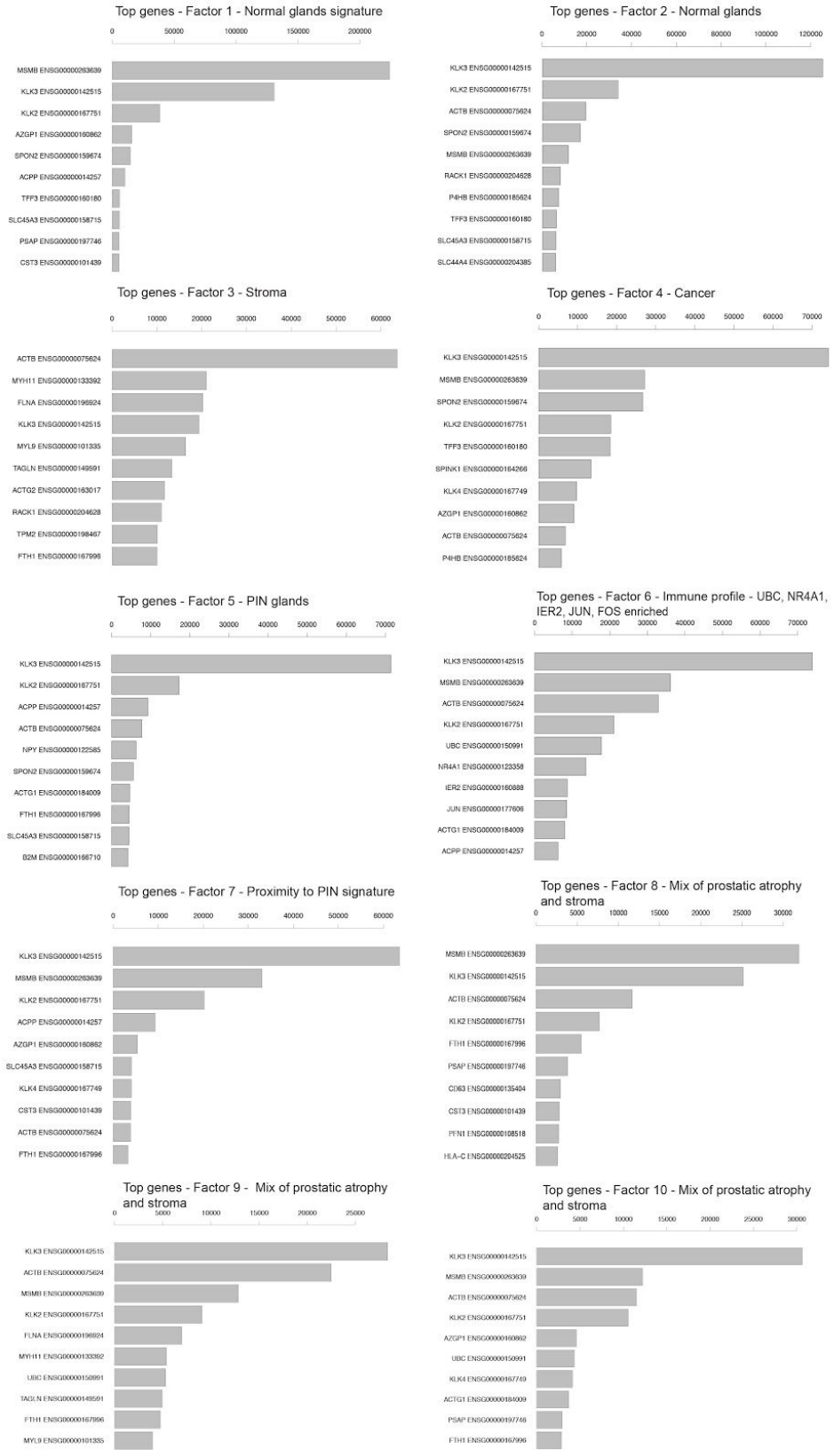
Supplementary Figure 1a. Spatially-resolved transcriptome profiling in prostate cancer. Tissue sections are placed on the array, fixed, H&E stained and imaged. The tissue is permeabilized, transcripts are captured and reverse transcribed. After tissue removal, barcoded cDNA is enzymatically released from the array and used for further library preparation and sequencing. The spatial barcode is used to connect every transcript with the spot it derives from, the UMI to correct for PCR amplification bias.



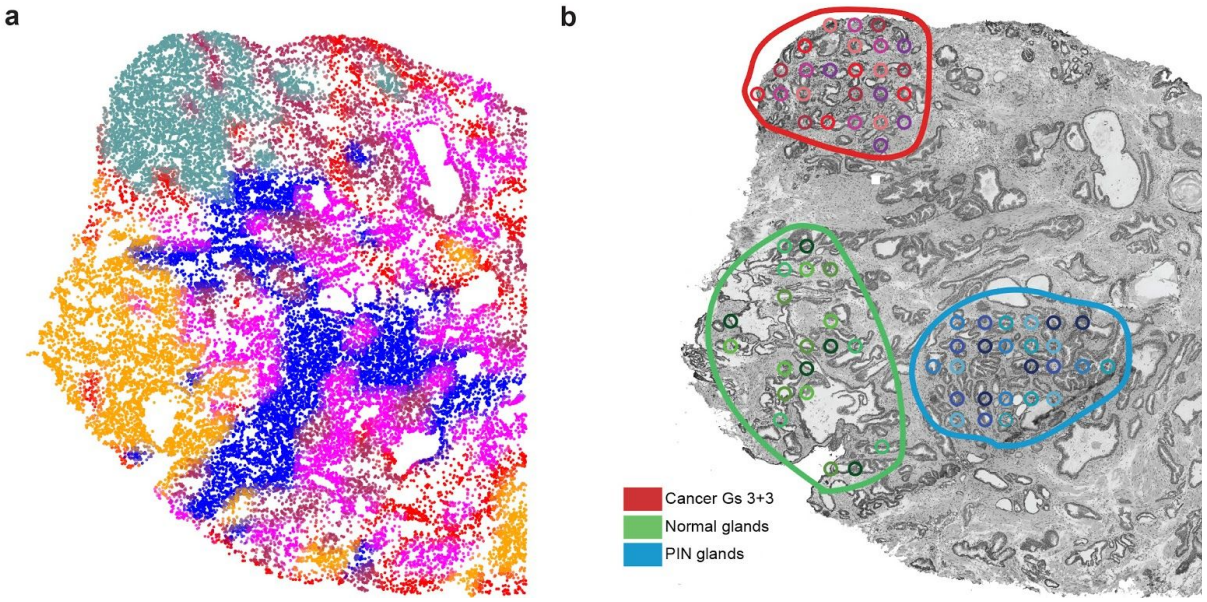
Supplementary Figure 1b. Histological prostate cancer tissue sections annotated by a pathologist are colored; cancer (red), PIN (yellow), suspected cancer (green) and inflammation (purple).



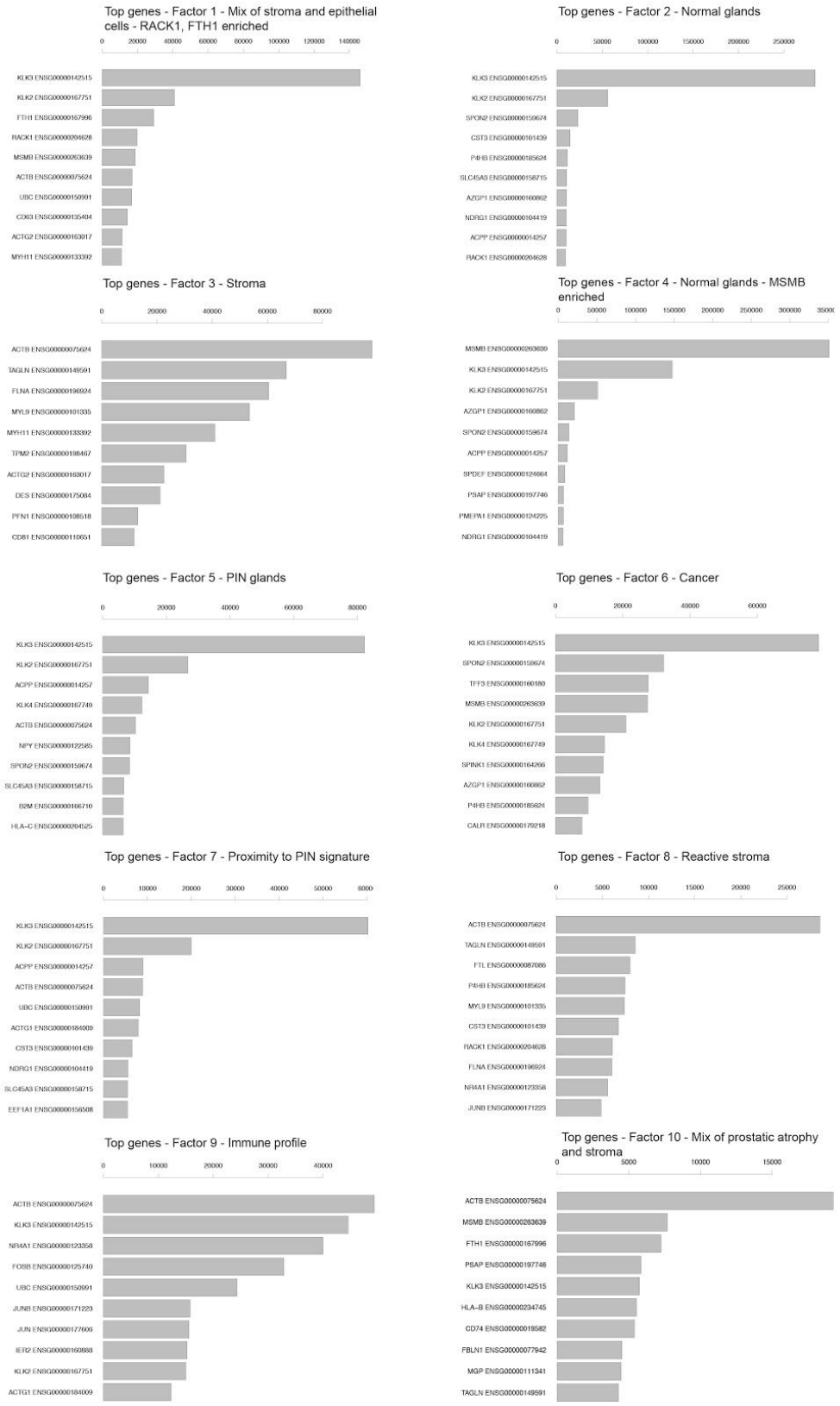
Supplementary Figure 2. Remaining activity maps from factor analysis in Fig. 2. a Factor activity maps of one cancer sample corresponding to normal gland signatures, proximity to PIN signature and mix of stroma and prostatic atrophy. Enlarged boxes show examples of the histology specific for the given factor. **b** Hierarchical clustering of all 10 factors.



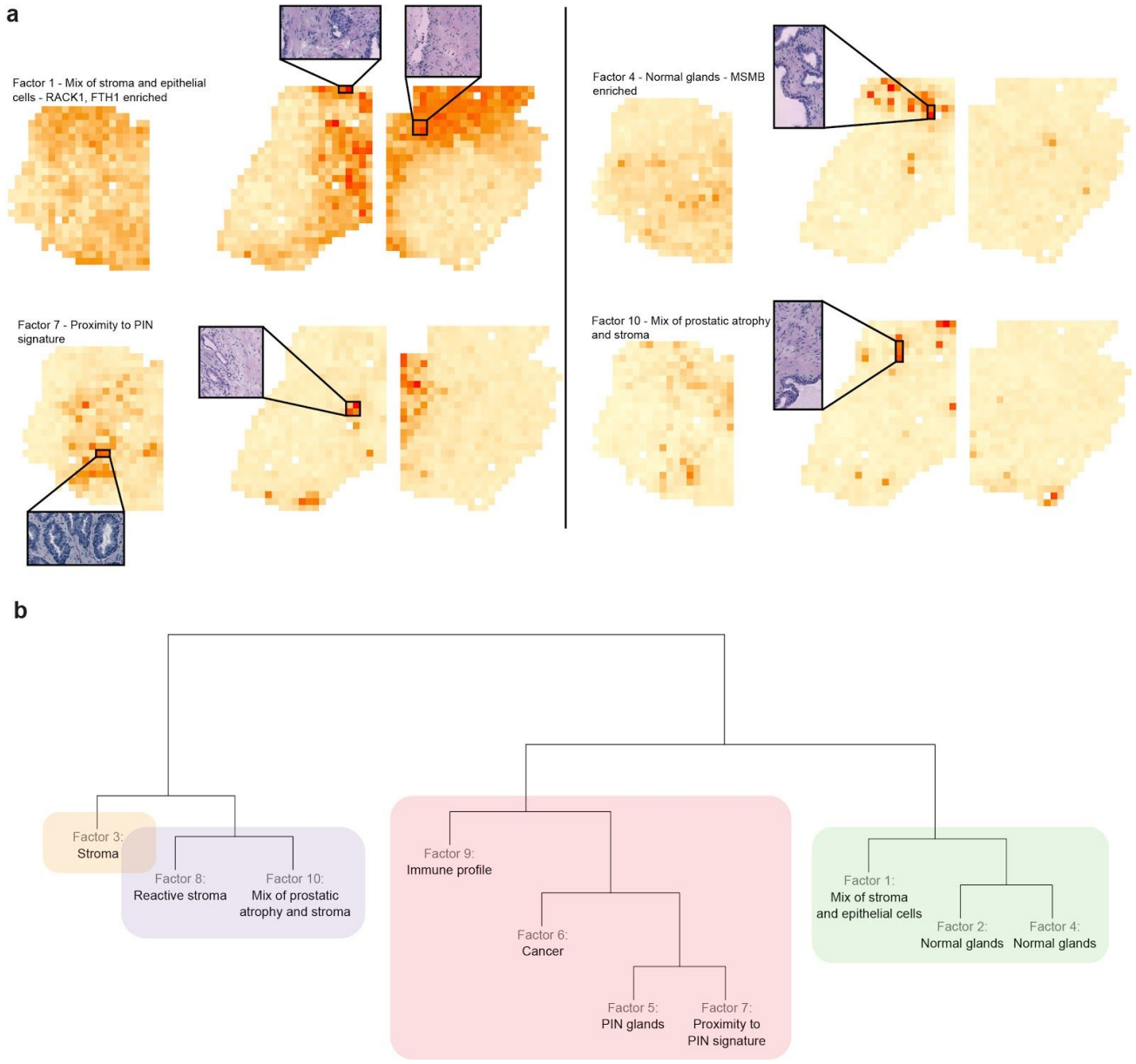
Supplementary Figure 3. Expected number of reads explained by all ten factors of the factor analysis in Fig. 2a. Only the ten highest-expressed genes are shown.



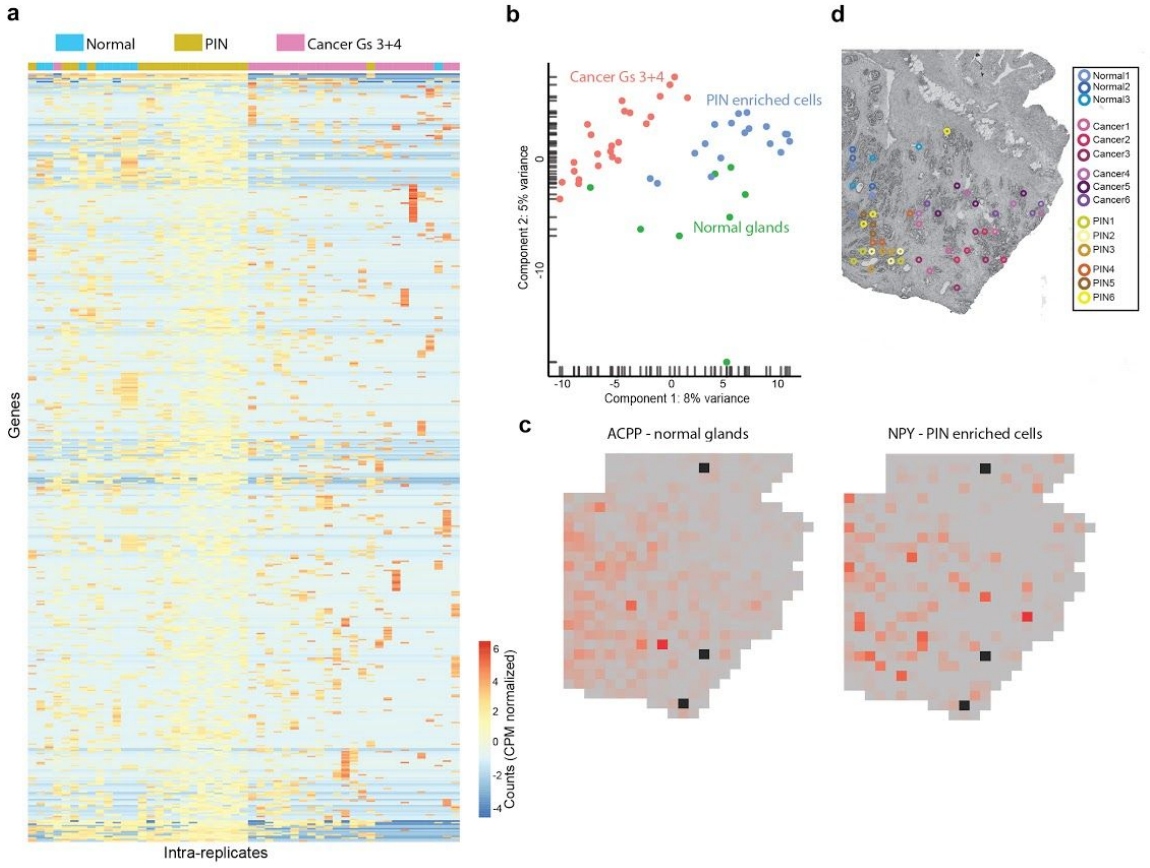
Supplementary Figure 4. a Hierarchical clustering of the spatial features revealed 6 clusters. Each cluster was assigned a color and the cluster identity was interpolated across the tissue structure to visualize major spatial patterns within the sample. This analysis correlates well with our factor based method and the morphology: orange: normal epithelium, green: cancer, blue and pink: PIN, red and purple: stroma. **b** We manually defined three regions based on factor activities and morphological information; Gs 3+3 (red), normal glands (green) and PIN (blue). In addition, spot replicates extracted from three different areas used for downstream analysis are shown. Each replicate (with 4-5 spots) is uniquely colored.



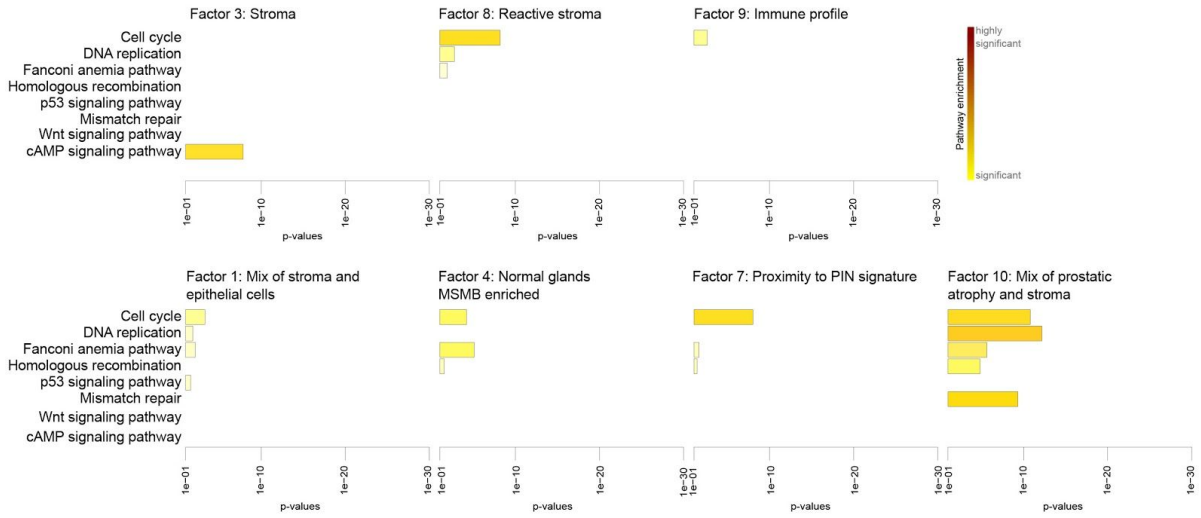
Supplementary Figure 5. Expected number of reads explained by all ten factors from factor analysis in Fig. 3b. Only the ten highest-expressed genes are shown.



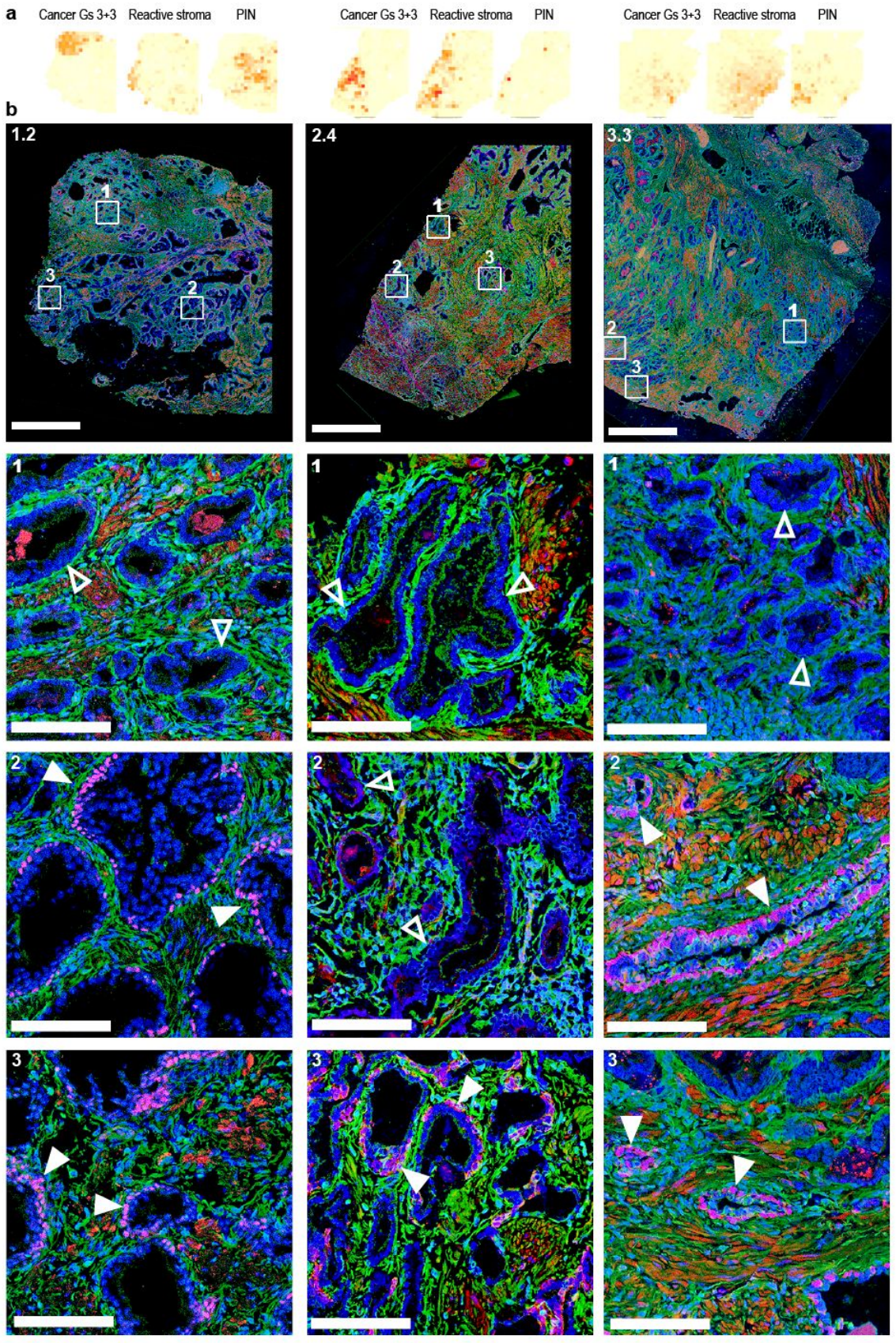
Supplementary Figure 6. Remaining activity maps from factor analysis in Fig. 3b. a Factor activity maps of three cancer containing samples corresponding to either normal glands or mix of stroma and epithelial cells. Enlarged boxes show examples of the histology specific for that factor. **b** Hierarchical clustering suggest that immune reactive stroma and stroma cluster close whereas PIN, cancer and inflammation are observed in another cluster. Normal glands (with and without MSMB) are neither similar to cancer nor to stroma cells.



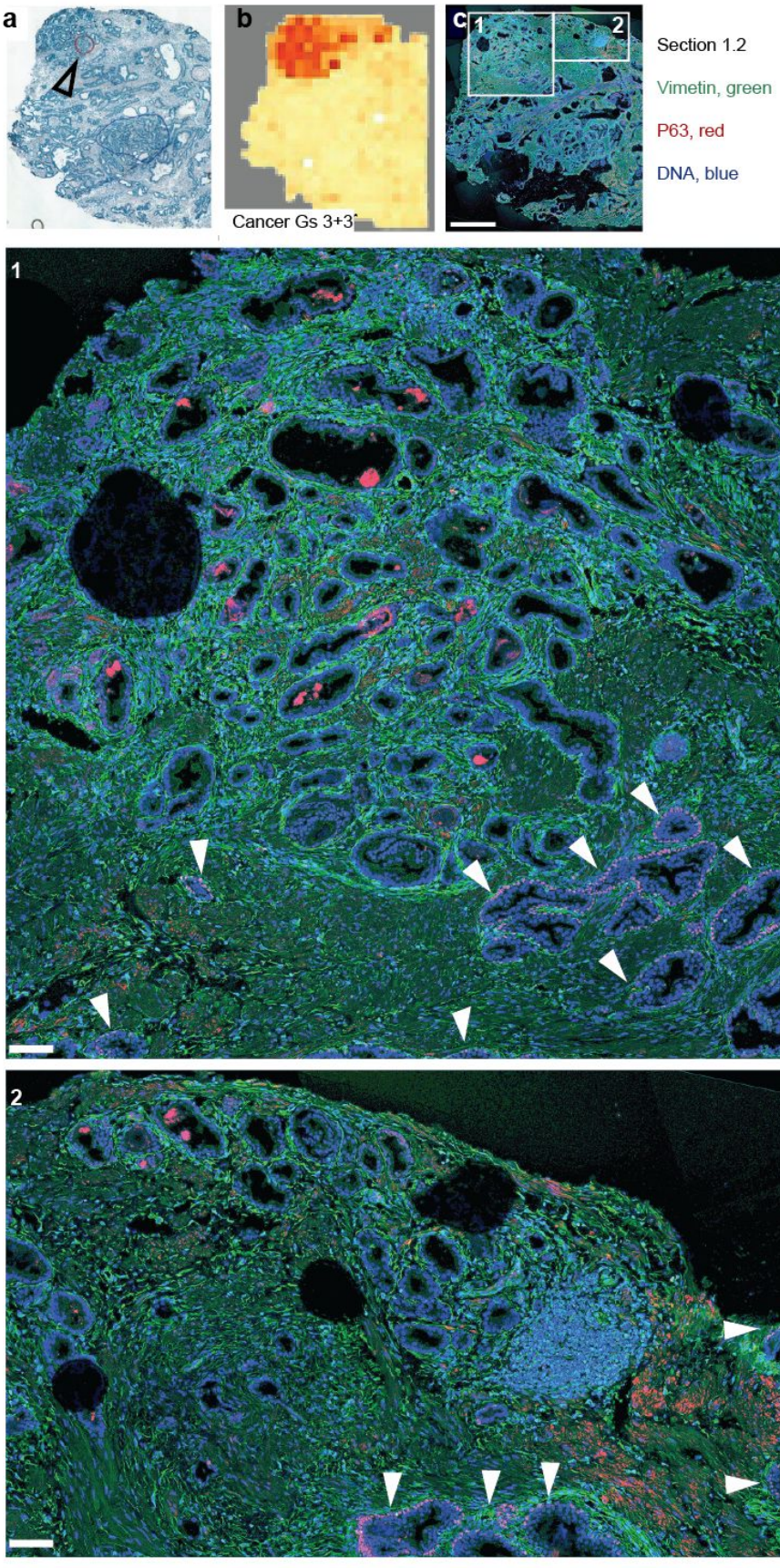
Supplementary Figure 7. Analysis between three factor activity clusters in sample 3.3 (cancer, normal and PIN enriched areas). **a** Heatmap of the 500 most variable genes between cancer, PIN and normal regions suggest that the normal and PIN regions identified by ST are different from the cancer region. **b** PCA plot separates normal and PIN glands from cancer cells. **c** Heatmaps of ACPP and NPY genes with absolute transcript counts show that the genes are expressed more in the PIN and normal regions compared to the cancer region identified by ST. **d** Spot replicates extracted from three different areas used for analysis in **a** and **b** are shown. Each replicate (with 3-4 spots) is uniquely colored.



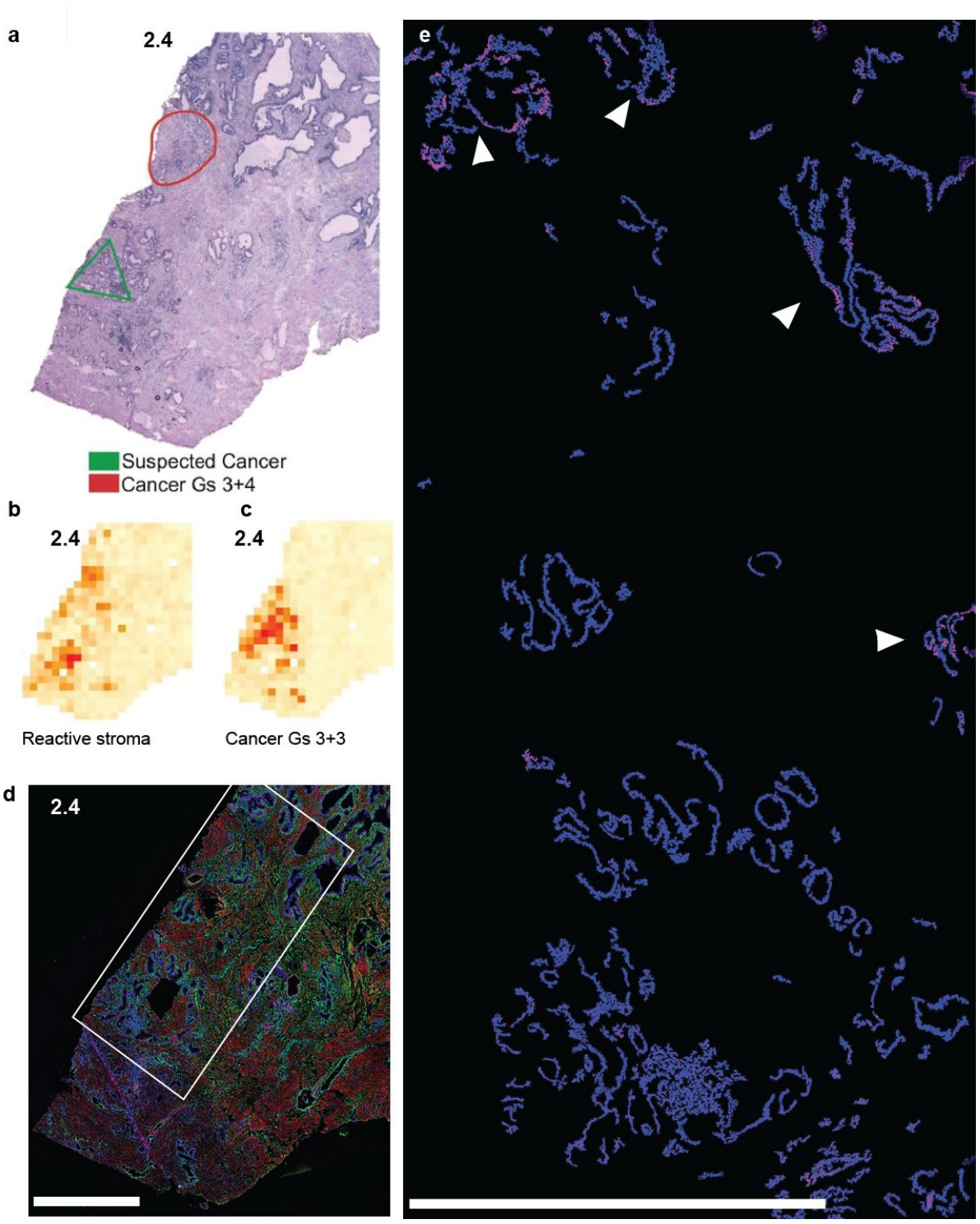
Supplementary Figure 8. Pathway enrichment. Remaining pathways enriched in all cancer samples from analysis seen in Fig. 3c.


a
Cancer Gs 3+3
Reactive stroma
PIN
Cancer Gs 3+3
Reactive stroma
PIN
Cancer Gs 3+3
Reactive stroma
PIN
b
1.2
2.4
3.3

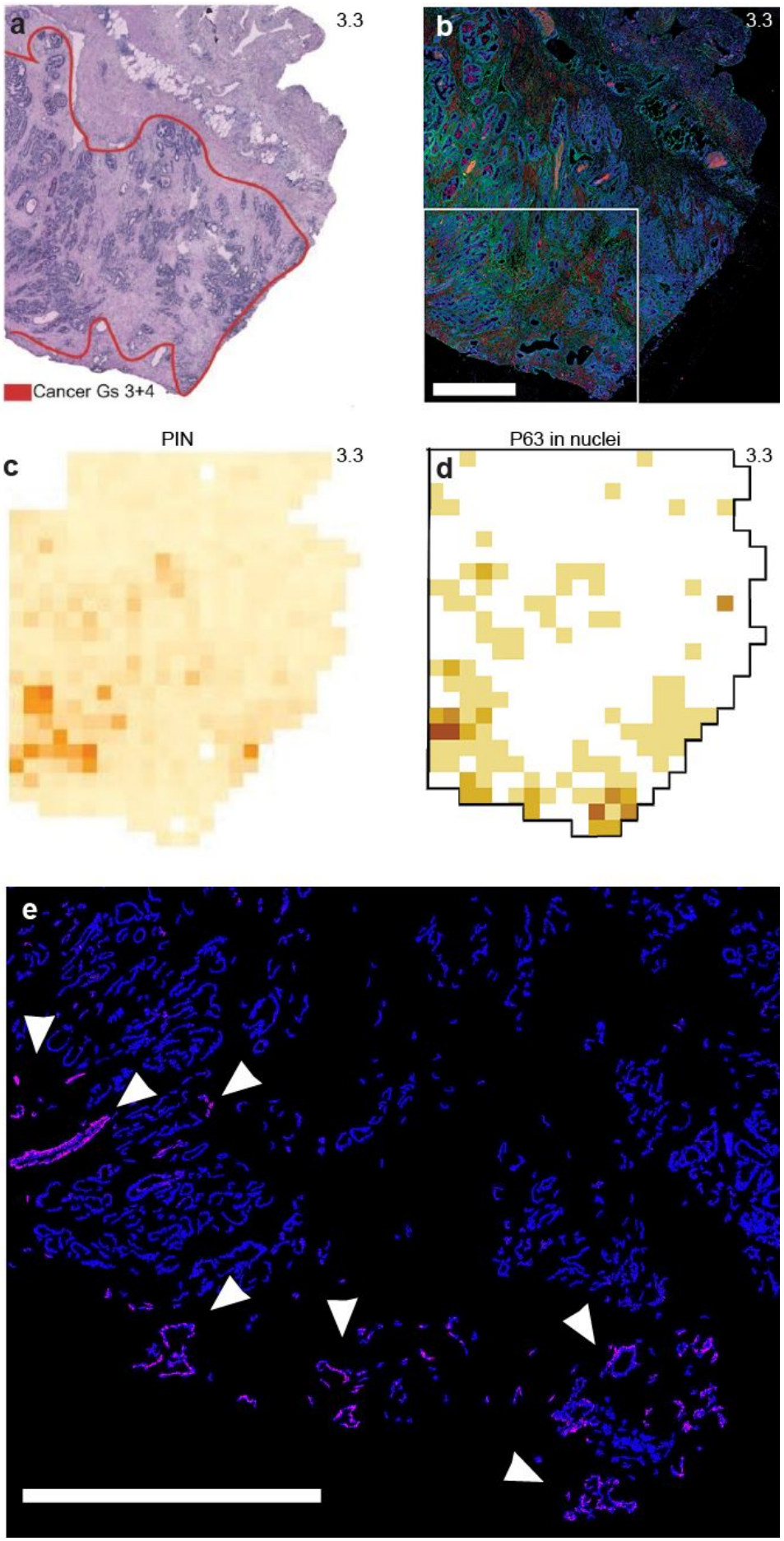
Supplementary Figure 9. Glands in cancer areas lack P63-staining basal cells. **a** Factor activity maps from the factor analysis of the three cancer samples reproduced from Fig. 3. **b** Fluorescence microscopy images of neighboring tissue sections stained by IHC for P63 (red), vimentin (green) and DNA (blue). Areas marked by white rectangles in the tissue sections are shown as close-ups below each section. Filled and unfilled triangles respectively point at glands with or without P63-stained basal cells. Scale bars in tissue sections and close-ups respectively indicate 1 mm and 100 μm


a
b
Cancer Gs 3+3
c
1
2
Section 1.2
Vimetin, green
P63, red
DNA, blue
1
2

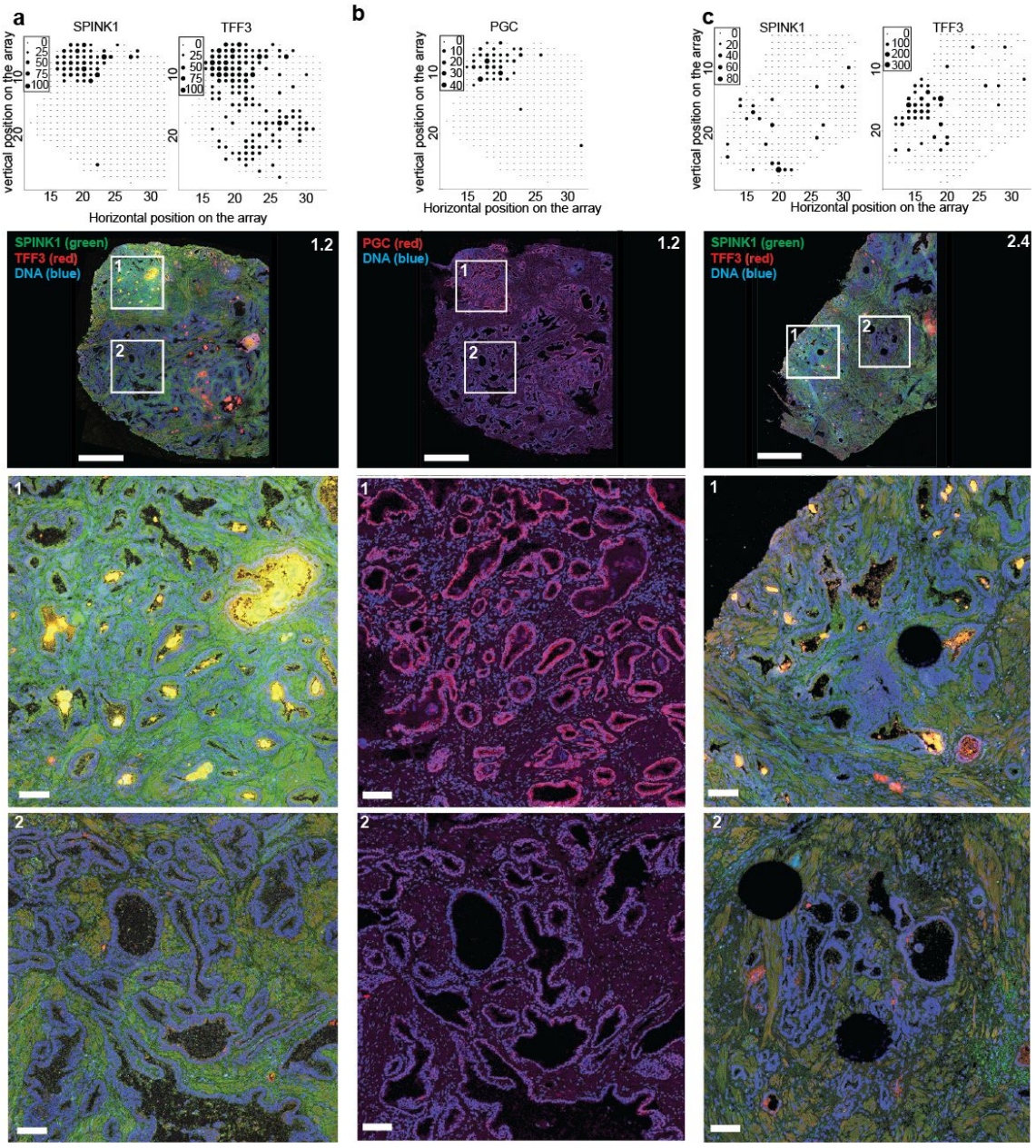
Supplementary Figure 10. Glands in the area marked as cancer by the ST "cancer" factor lack P63 staining. a H&E staining of section 1.2 with the cancer annotation done by the pathologist. **b** Factor activity maps from the factor analysis of the three cancer samples reproduced from Fig. 3. **c** Fluorescence microscopy images of the tissue sections stained by IHC for vimentin (green), P63 (red) and DNA (blue). Areas marked with white rectangles in the tissue sections corresponds to, and are slightly bigger than the "cancer" factor in **b**. Close up of the white rectangles are shown in the middle and lower image. Filled triangles point at glands with P63-stained basal cells. Notice that glands outside the area of the "cancer" factor have basal cells with P63 staining. Scale bars in tissue sections and close-ups respectively indicate 1 mm and 100 μm.


a
2.4
Suspected Cancer
Cancer Gs 3+4
b
2.4
Reactive stroma
c
2.4
Cancer Gs 3+3
d
2.4
e

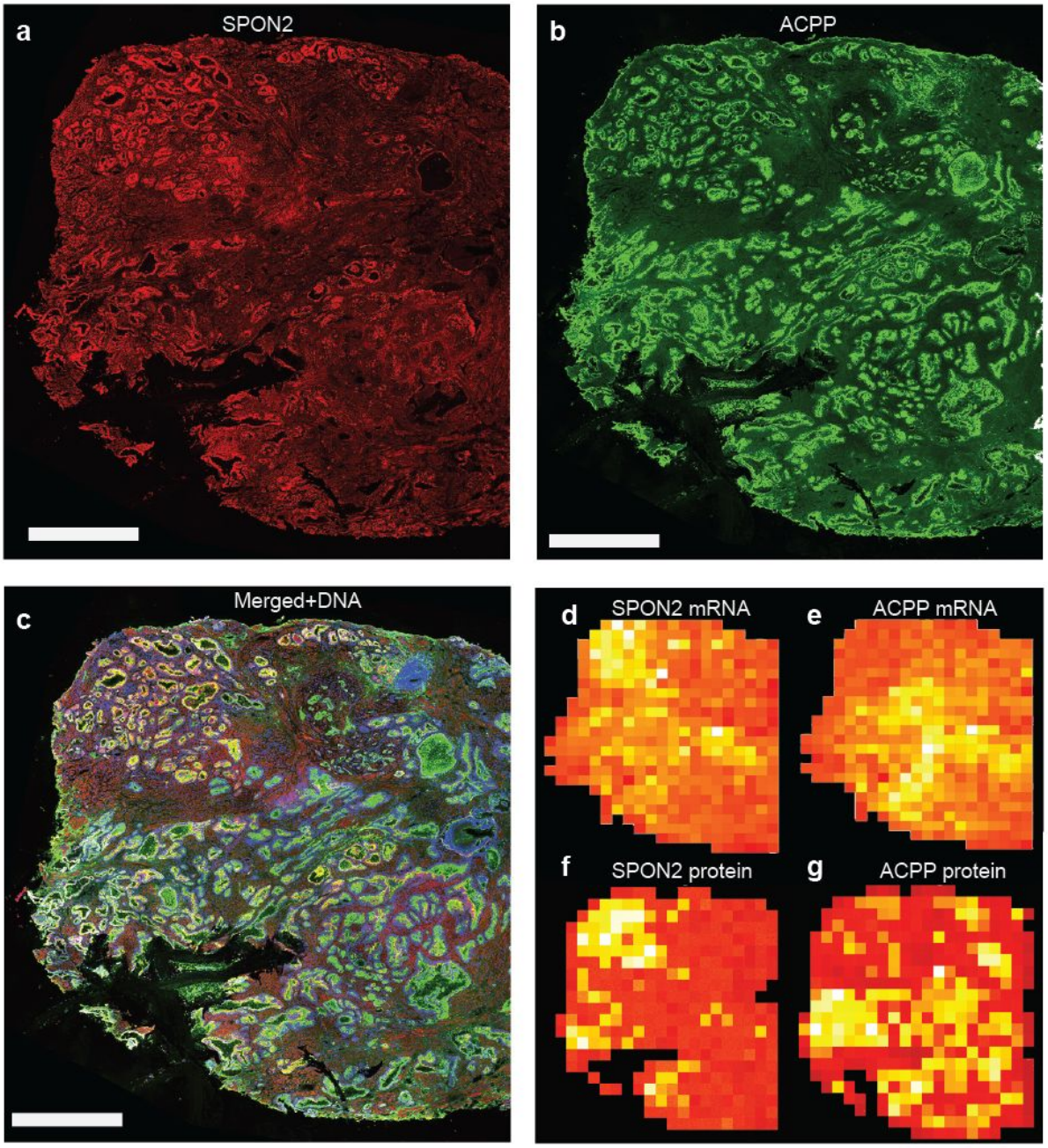
Supplementary Figure 11. Glands in the area marked as cancer by the ST "cancer" factor lack P63 staining. a H&E staining of section 2.4 with the cancer annotation done by the pathologist. **b and c** Factor activity maps from sample 2.4 reproduced from Fig. 3. **d** Fluorescence microscopy images of the tissue sections stained by IHC for vimentin (green), P63 (red) and DNA (blue). Area marked with white rectangle in the tissue section corresponds to and is slightly larger than the "cancer" and "reactive stroma" factors in **b** and **c**. **e** Close up of box in **d**. The P63 signal shown is extracted by using the DNA image as a mask. Only P63 signal from epithelial nuclei is shown. Filled triangles point at glands with P63-stained basal cells. Notice that glands outside the area of the cancer factor in the upper part of the close up have basal cells with P63 staining. Scale bars in tissue sections and close-ups respectively indicate 1 mm.


a
3.3
Cancer Gs 3+4
b
3.3
PIN
c
3.3
P63 in nuclei
d
3.3
e

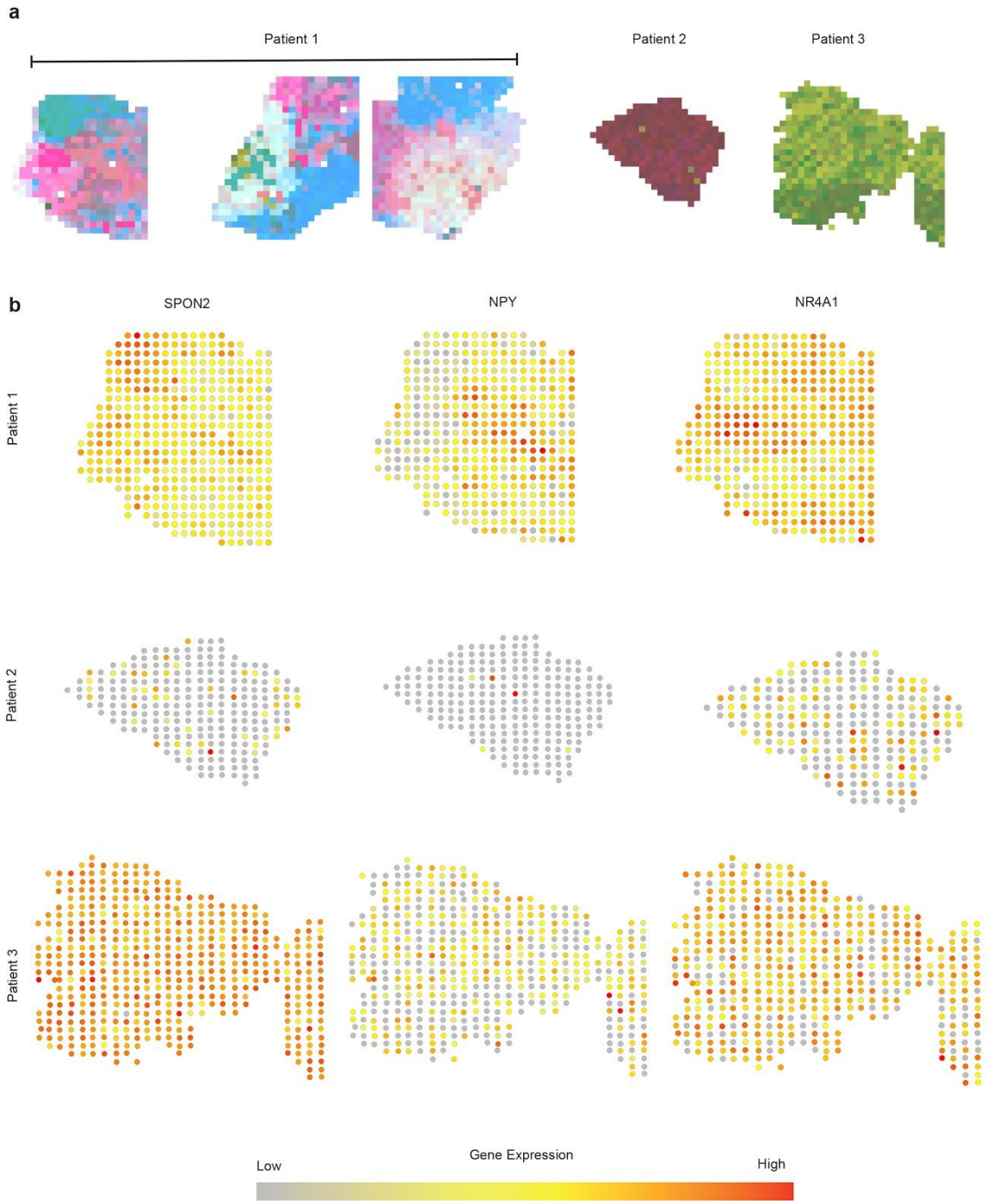
Supplementary Figure 12. Sample 3.3 shows activity of the ST "PIN" factor in a region annotated as cancer. IHC reveals that a fraction of glands in this region are P63 positive. **a** H&E staining of section 3.3 with the cancer annotation done by the pathologist. **b** Factor activity map for "PIN" from cancer sample 3.3 reproduced from Fig. 3. **c** Fluorescence microscopy images of the tissue sections stained by IHC for vimentin (green), P63 (red) and DNA (blue). Area marked with white rectangle in the tissue sections contains the area positive for the "PIN" factor in **b**. **d** Protein distribution map of P63 signal solely from epithelial nuclei. **e** Close up of box in **c**. Only P63 signal from epithelial nuclei is shown. Filled triangles point at glands with P63-stained basal cells. Note that the areas with basal cells correspond to the areas positive for the "PIN" factor and areas annotated as cancerous or occasional PIN glands by the pathologist. Scale bars in tissue sections and close-ups respectively indicate 1 mm**.**


a
SPINK1
TFF3
vertical position on the array
Horizontal position on the array
b
PGC
vertical position on the array
Horizontal position on the array
c
SPINK1
TFF3
vertical position on the array
Horizontal position on the array
SPINK1 (green)
TFF3 (red)
DNA (blue)
1.2
PGC (red)
DNA (blue)
1.2
SPINK1 (green)
TFF3 (red)
DNA (blue)
2.4
1
2

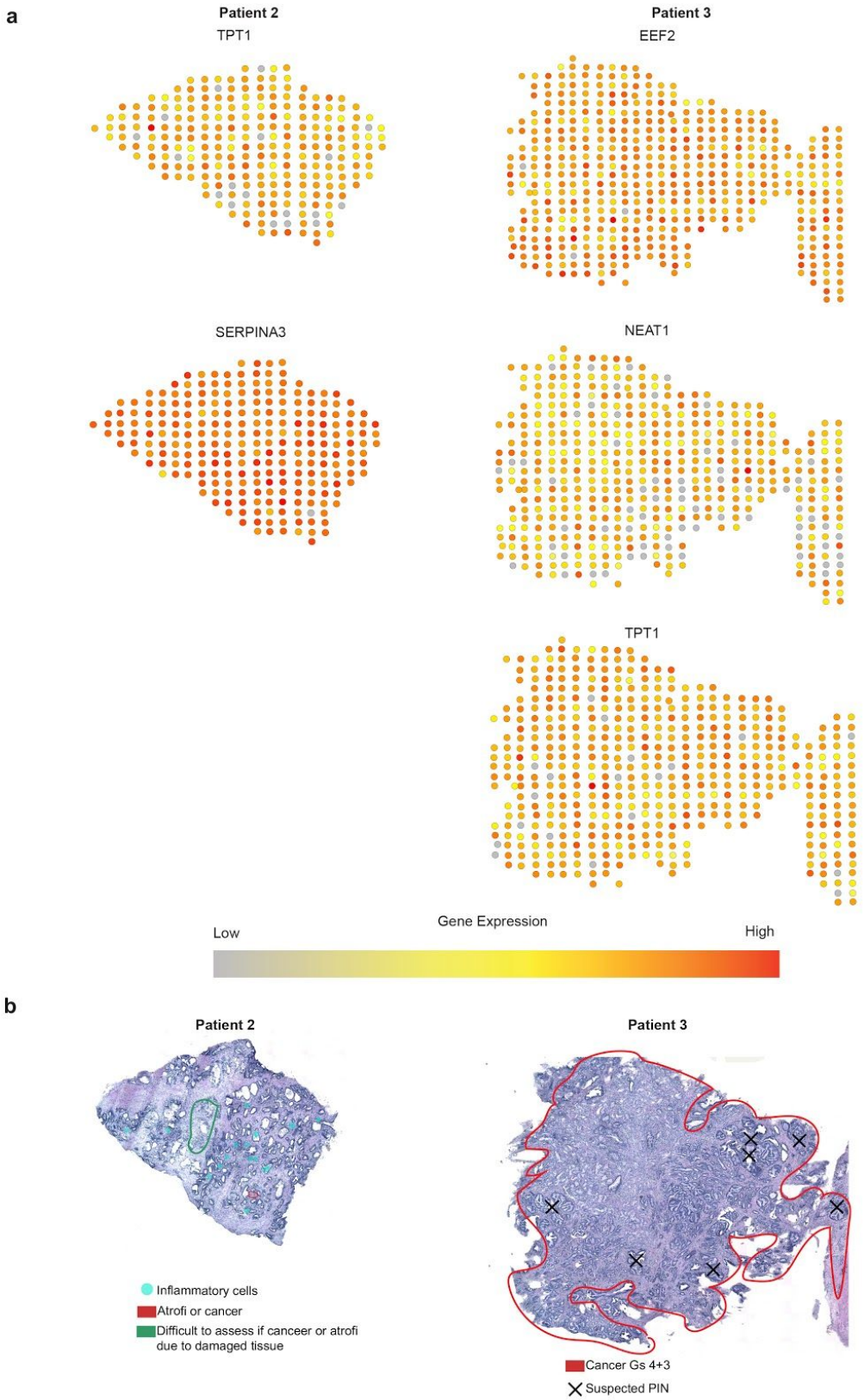
Supplementary Figure 13. Consistent spatial expression patterns of mRNA and protein measured by ST and IHC. Circle size in array dot plots indicates normalized ST counts. Proteins stained by IHC indicated by colored labels, nuclei stained with DAPI (blue). Areas marked with numbered, white rectangles in the tissue sections contain cancerous (1) and normal regions (2) and are shown as close-ups in the two bottom rows. Scale bars in tissue sections and close-ups indicate 1mm and 100µm, respectively. **a** SPINK1 and TFF3 in section 1.2. **b** PGC in section 1.2. **c** SPINK1 and TFF3 in section 2.4. Note: colocalization of SPINK1 and TFF3 shows as yellow. SPINK1, TFF3, PGC expression levels were zero in the ST data for section 3.3. and hence were excluded from IHC.



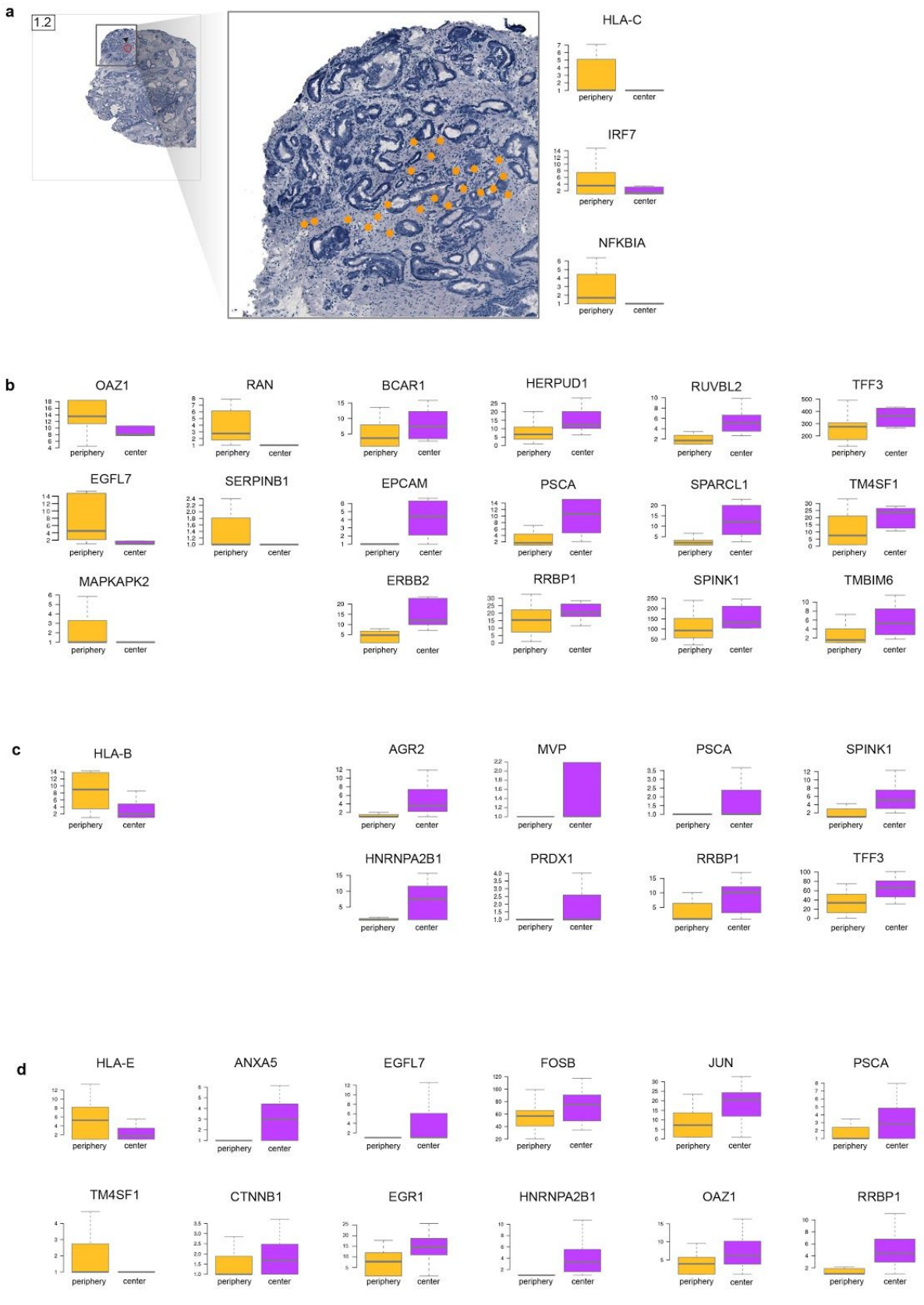
Supplementary Figure 14. Comparison of protein and mRNA localisation in tissue section 1.2. a IHC image of SPON2 (red). **b** IHC image of ACPP (green). **c** Merged image of SPON2, ACPP and DNA (blue). **d** Heatmap of ST-mRNA data for SPON2. **e** Heatmap of ST-mRNA data for ACPP. **f** Heatmap of IHC data in **a**. **g** Heatmap of IHC data in **b**. Scale bars in IHC images indicate 1 mm.



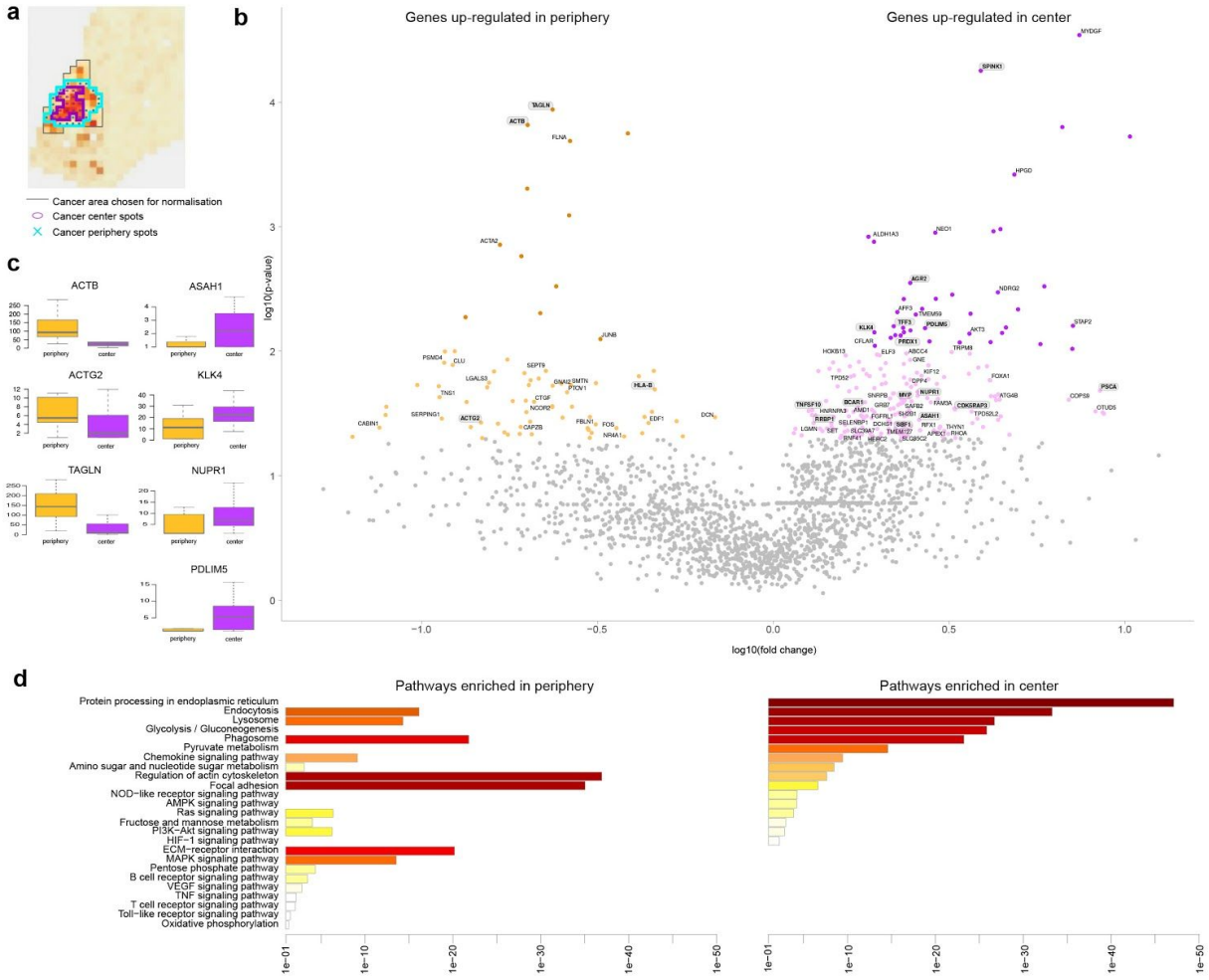
Supplementary Figure 15. **Analysis between 3 patients.** Results demonstrate high inter-tumor heterogeneity, although some genes (e.g. SPON2, NPY and NR4A1) in patient 1 were also shared between patient 2 and 3.



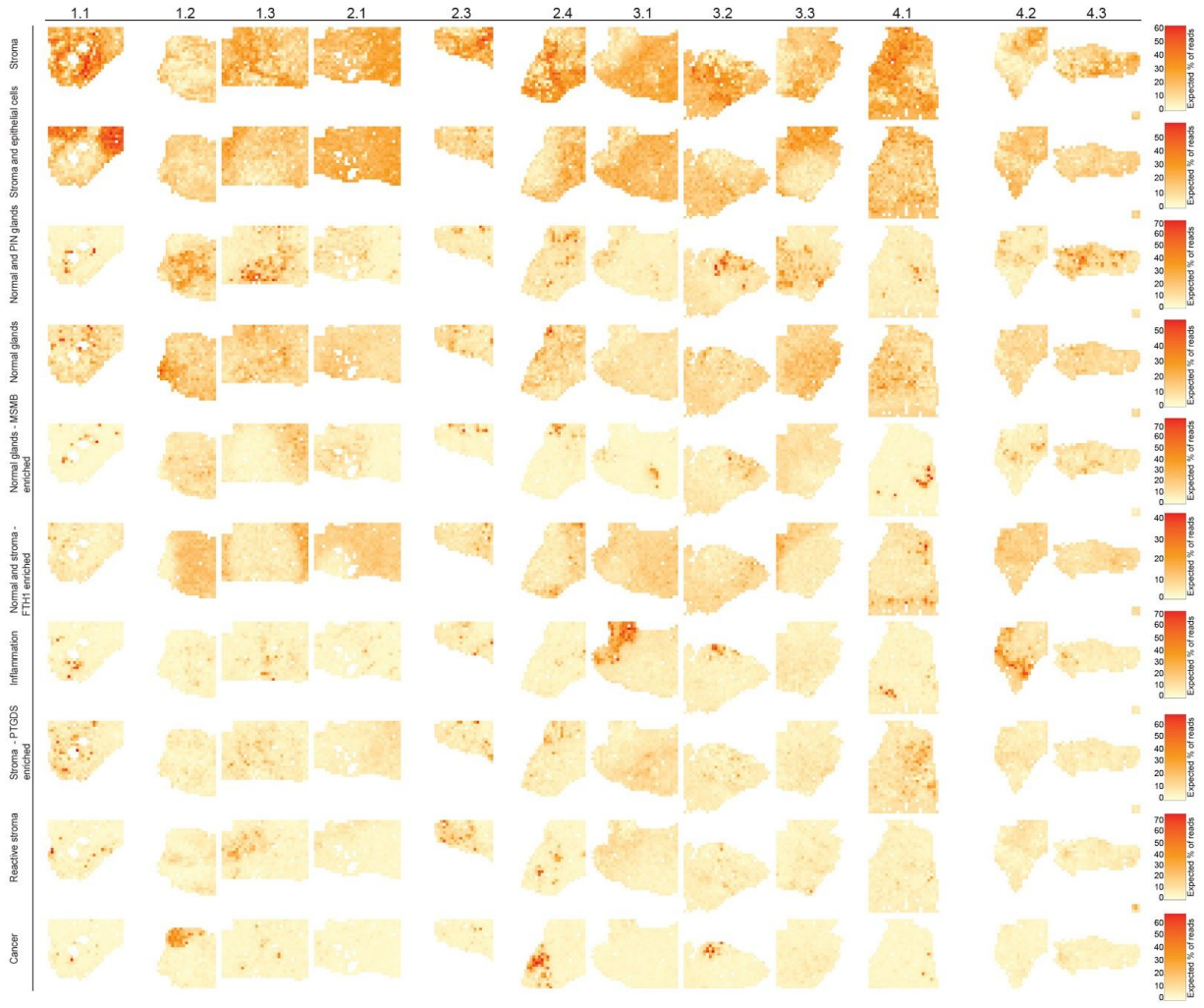
Supplementary Figure 16. Examples of unique sets of transcripts in each tumor. a TPT1 and SERPINA3 was highly expressed in patient 2 and EEF2, NEAT1 and TPT1 in patient 3. **b** Histological prostate cancer tissue sections annotated by a pathologist are colored.


a
1.2
HLA-C
IRF7
NFKBIA
b
OAZ1
RAN
BCAR1
HERPUD1
RUVBL2
TFF3
EGFL7
SERPINB1
EPCAM
PSCA
SPARCL1
TM4SF1
MAPKAPK2
ERBB2
RRBP1
SPINK1
TMBIM6
c
HLA-B
AGR2
MVP
PSCA
SPINK1
HNRNPA2B1
PRDX1
RRBP1
TFF3
d
HLA-E
ANXA5
EGFL7
FOSB
JUN
PSCA
TM4SF1
CTNNB1
EGR1
HNRNPA2B1
OAZ1
RRBP1
periphery
center

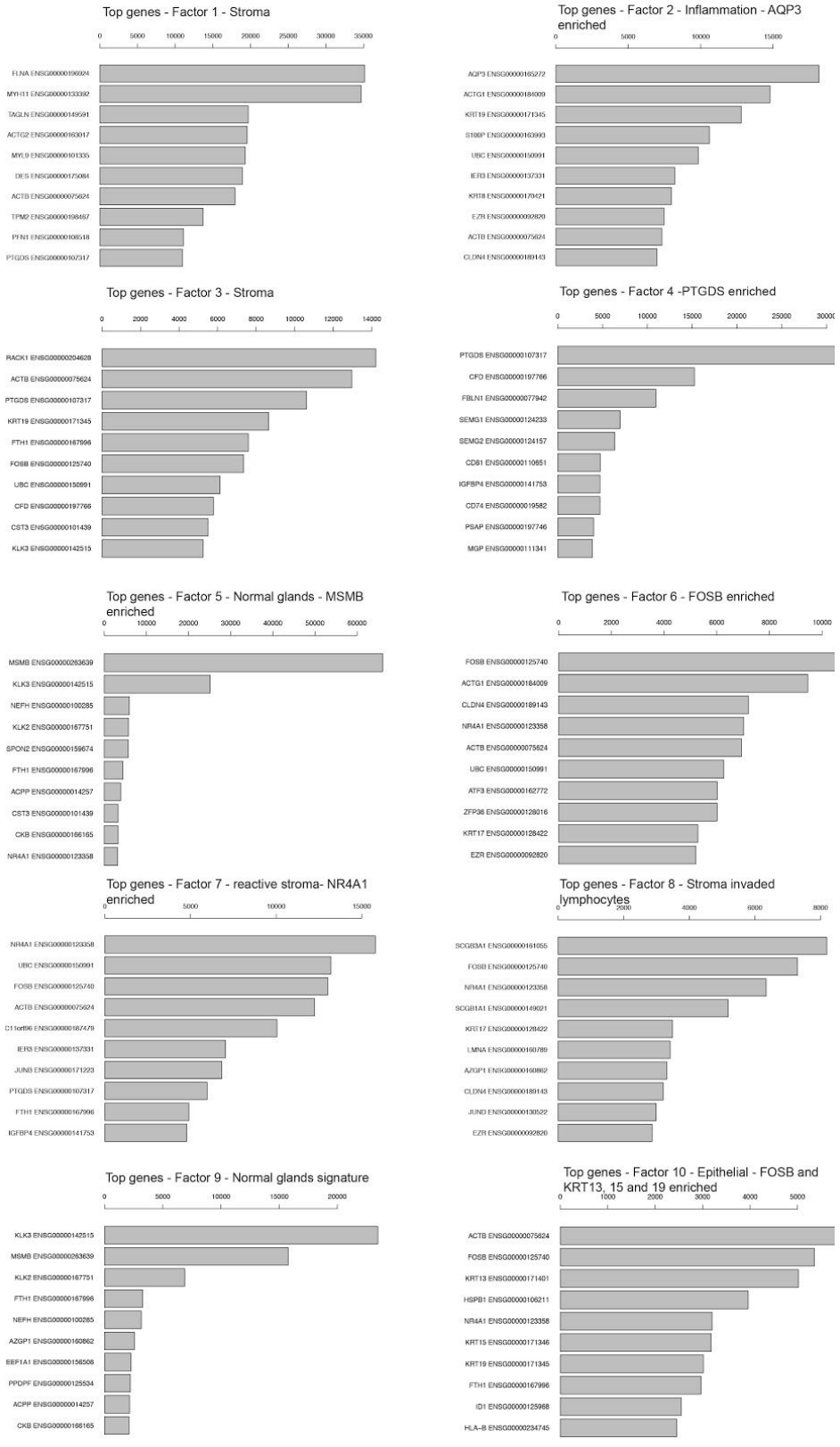
Supplementary Figure 17. a Annotated inflammation (orange color) by a pathologist and immune-related genes upregulated in the periphery of sample 1.2. **b, c, d** Remaining box plots from analysis in Fig. 4 showing expression levels of noteworthy genes significantly upregulated in either periphery or the cancer center of sample 1.2 (b), sample 2.4 (c) and sample 3.3 (d).



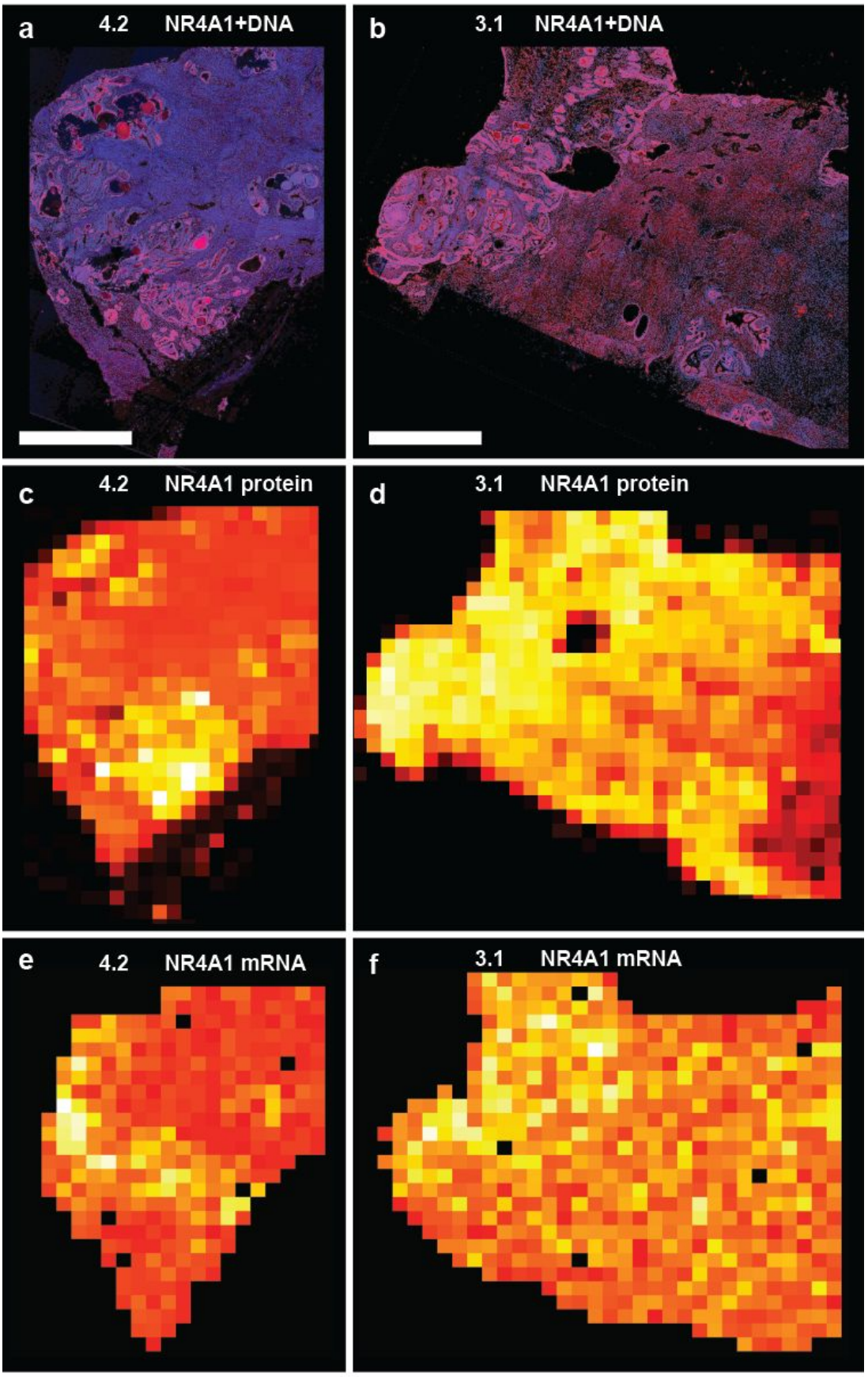
Supplementary Figure 18. Spatial comparison of periphery and center of sample 2.4. a Area comprising spots taken for normalisation of ST counts, within this area spots are chosen as periphery and center. Choice of spots is based on the activity of the factors “cancer” and “reactive stroma”. **b** Volcano plot of significantly differentially expressed genes between periphery and center. **c** Box plots showing expression levels of noteworthy genes significantly upregulated in either periphery or the cancer center. **d** Enriched pathways for significantly ($p < 0.05$) differentially expressed genes in center and periphery.



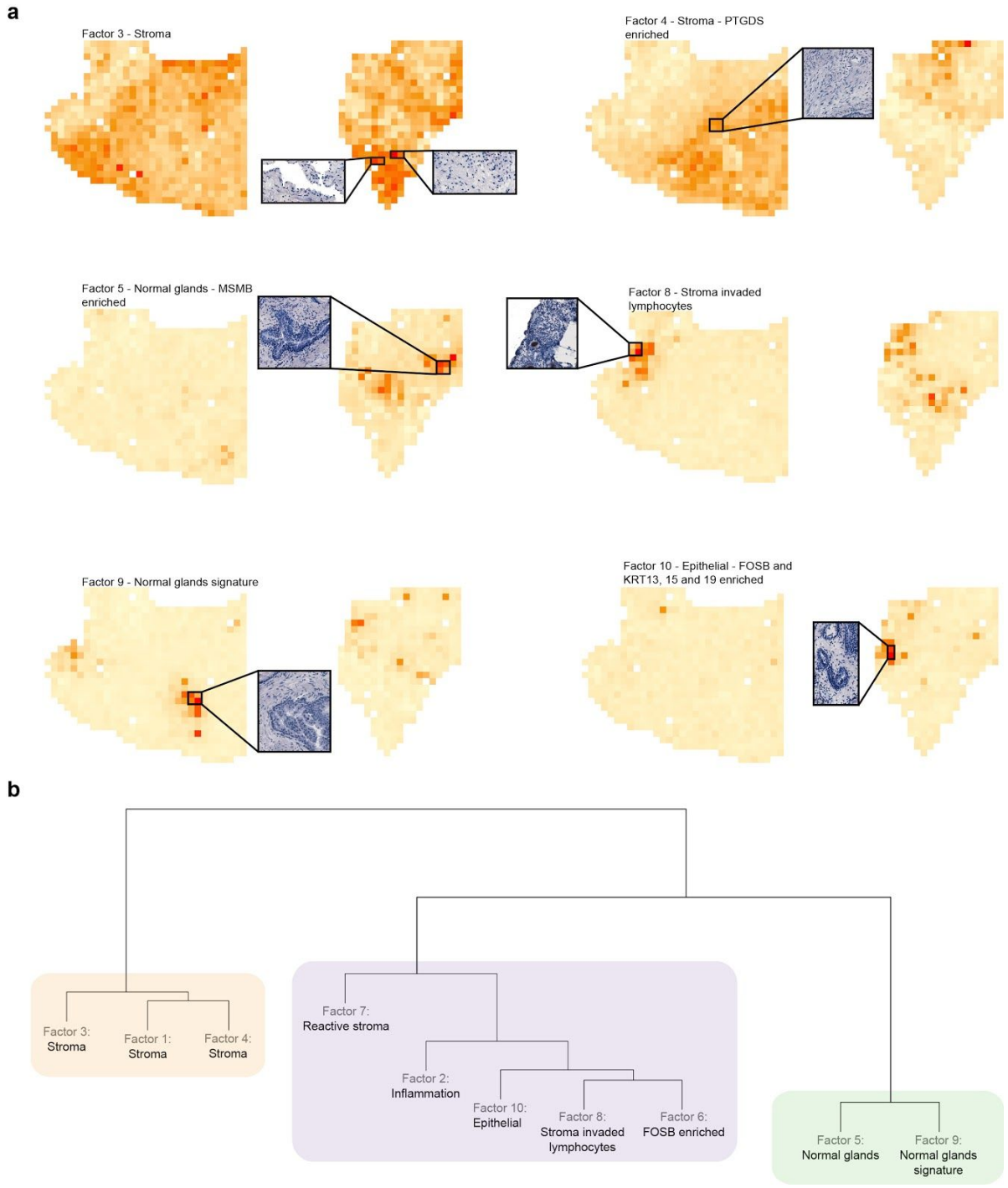
Supplementary Figure 19. Factor activity maps based on an analysis of all 12 samples. Scale bars equal relative frequencies expressed in percentage.



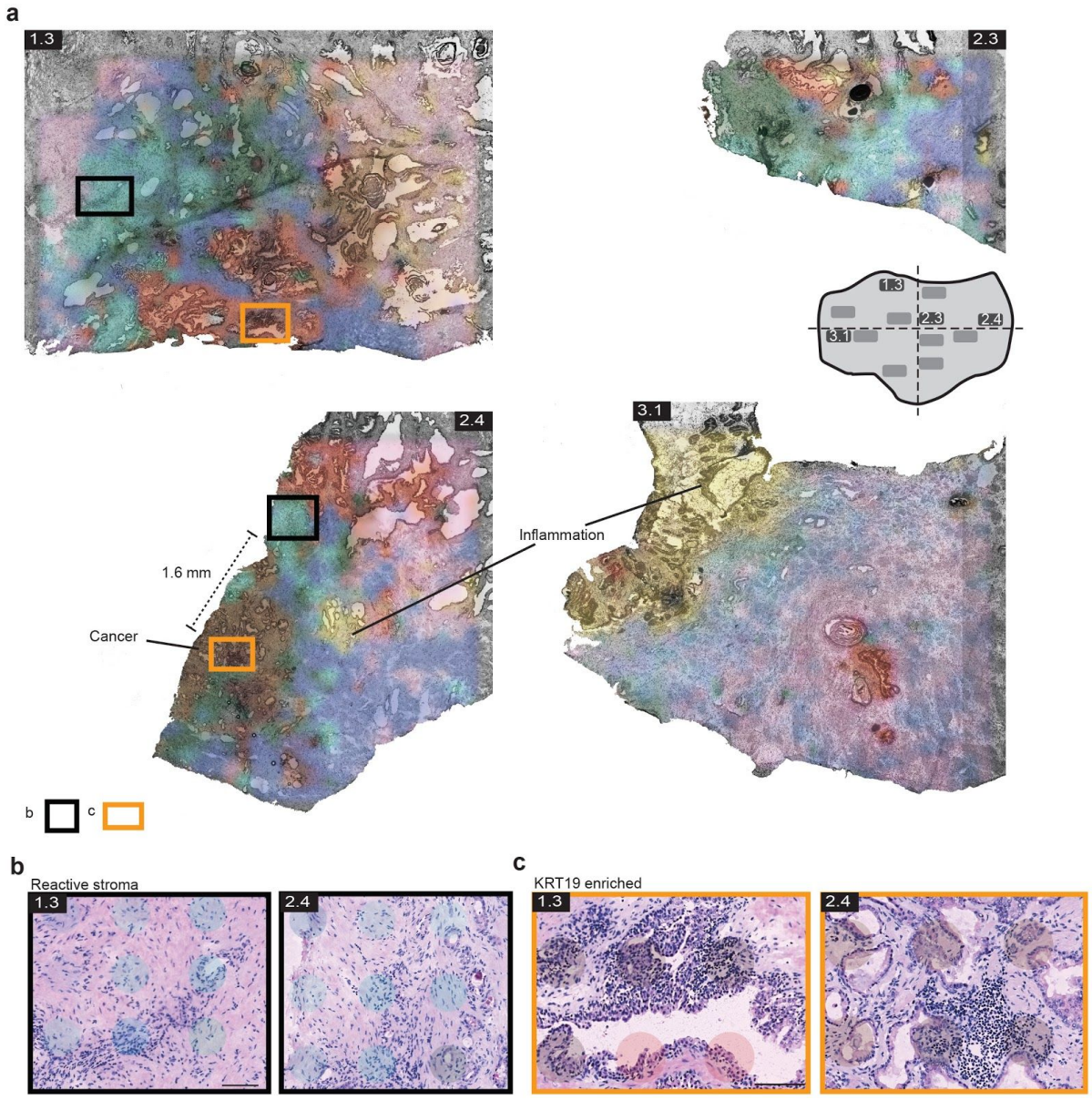
Supplementary Figure 20. Expected number of reads explained by factors from factor analysis in Fig. 5b. Only the ten highest-expressed genes are shown.


a
4.2
NR4A1+DNA
b
3.1
NR4A1+DNA
c
4.2
NR4A1 protein
d
3.1
NR4A1 protein
e
4.2
NR4A1 mRNA
f
3.1
NR4A1 mRNA

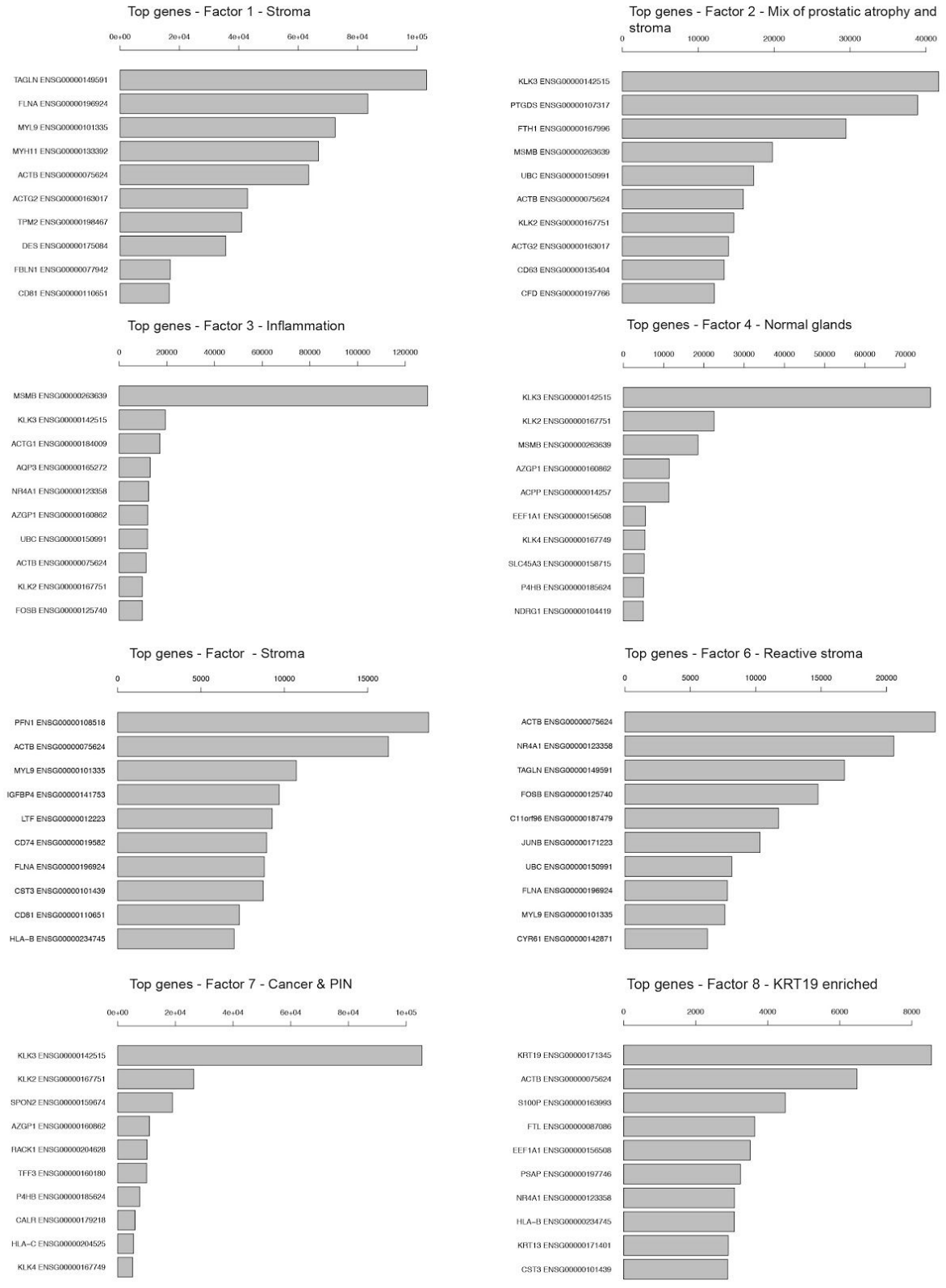
Supplementary Figure 21. Comparison of protein and mRNA localisation in tissue sections 4.2 and 3.1. a IHC image of NR4A1 (red) and DNA (blue) in section 4.2. **b** IHC image of NR4A1 (red) and DNA (blue) in section 3.1. **c** Heatmap of IHC data of NR4A1 shown in **a**. **d** Heatmap of IHC data of NR4A1 shown in **b**. **e** Heatmap of ST-mRNA data of NR4A1 in section 4.2. **f** Heatmap of ST-mRNA data of NR4A1 in section 3.1. Scale bars in IHC images indicate 1 mm.



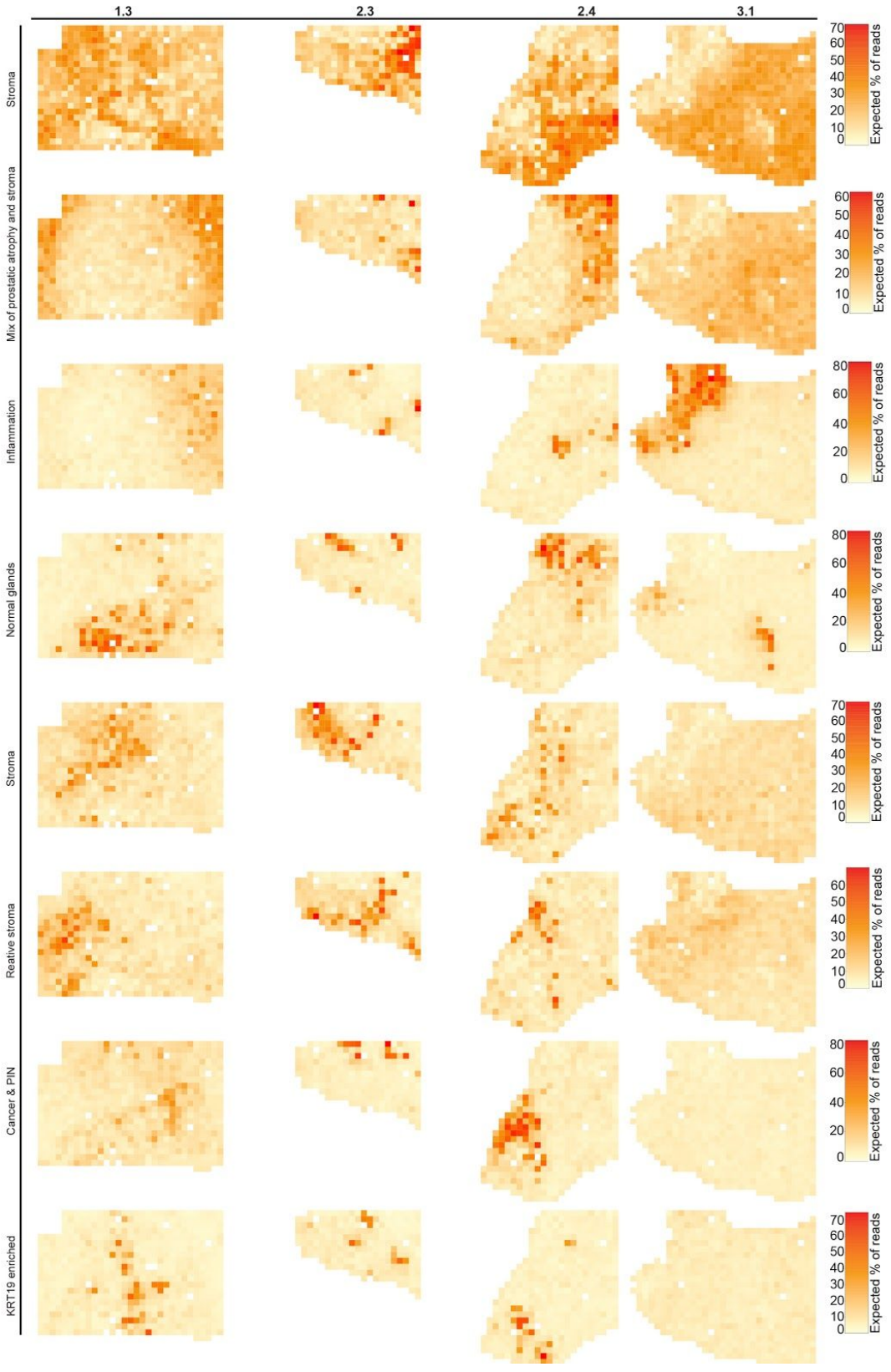
Supplementary Figure 22. Remaining activity maps from factor analysis in Fig. 5. a Factor activity maps of two samples with inflammation corresponding to different types of stroma or epithelial signatures. Enlarged boxes show examples of the histology specific for that factor. **b** Hierarchical clustering reveal that stroma cluster by their own, separated from normal, inflammation and reactive stroma.



Supplementary Figure 23. Microenvironment vicinal to cancer and inflammation. a t-SNE summary of factor activity maps based on a joint factor analysis of the four samples (Supplementary data 5), linearly interpolated and superimposed on histological images for visibility of morphology. **b and c** Close-ups of reactive stroma and KRT19-enriched regions shown in black and orange rectangles in **a**. Position of array spots are visible behind the tissue section.

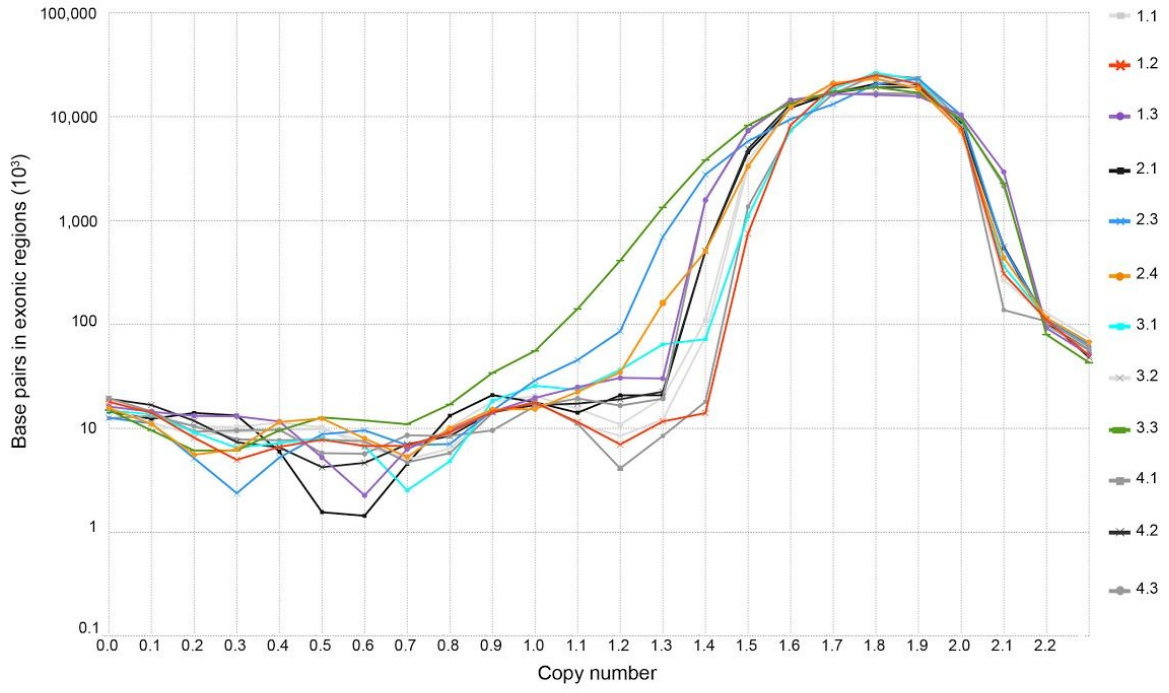


Supplementary Figure 24. Expected number of reads explained by factors from factor analysis in Supplementary Fig. 23. Only the ten highest-expressed genes are shown.

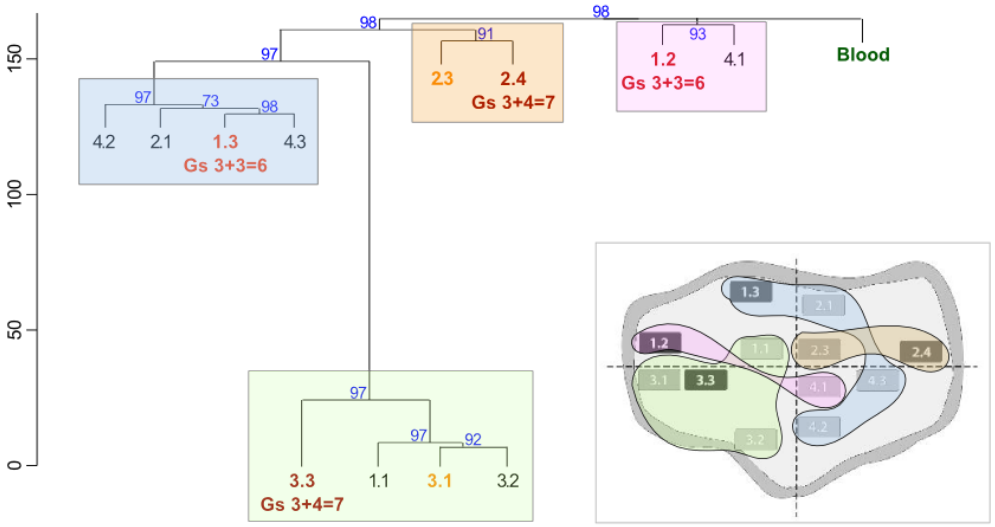


Supplementary Figure 25. Factor activity maps from analysis in Supplementary Fig. 23.

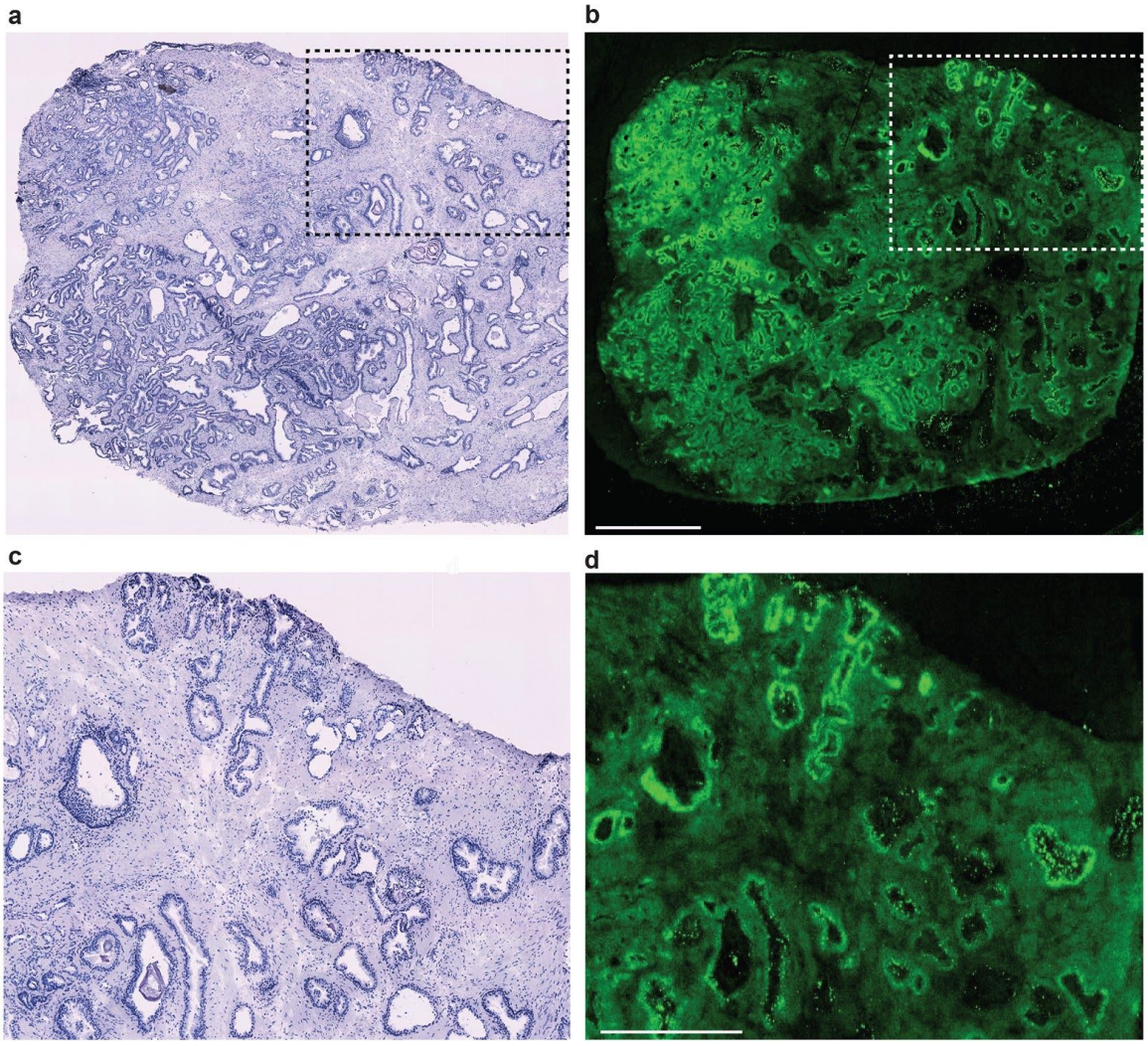
Scale bar equals relative frequencies expressed in percentage.



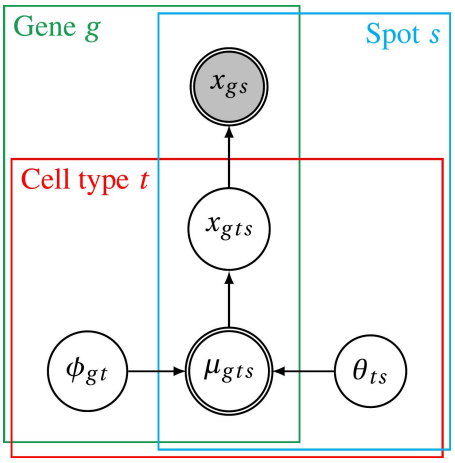
Supplementary Figure 26. Copy number analysis for deleted segments. Affected base pairs in exonic regions per rounded copy number value below 2.2 for each sample are presented. Sample 3.3 with the largest cancerous area shows the most deleted base pairs.



Supplementary Figure 27. **Similarity tree based on Euclidean distance and hierarchical clustering (ward.D2).** Segments of the whole genome with a CNV<1.6 (deletions) and >2.3 (amplifications) were considered. Four clusters were revealed and each cluster contains one cancerous sample. 1.2 was ending up close to blood in the tree, whereas sample 3.3 shows the biggest difference to blood. Further, 3.3 contains the highest number of genetic structural variations of the twelve samples.



Supplementary Figure 28. Quality control assay of a prostate cancer tissue section. a Hematoxylin and eosin staining of prostate cancer tissue. **b** Fluorescent cDNA signal after tissue removal. Scale bar=1.5mm. **c** and **d** Magnification of box in **a and b**. Scale bar=750µm. The results display that a permeabilization time of 10 min gives maximized signal within tissue and minimized diffusion outside the tissue.



Supplementary Figure 29. Graphical model representation of the core Poisson regression model. Plates for genes, spots, and cell types indicate replication. Circled nodes are random variables. Doubly-circled nodes are deterministic random variables. The shaded node is observed.

Supplementary Tables

Supplementary Table 1. Overview of samples and data evaluation for patient 1. Table includes pathological annotation and sequencing data statistics for all tissue sections, number of spots covered by tissue, genes per spots, unique transcripts covered by tissue and unique transcripts per spot. Also, the (relative) area in % of the given cell type is given. Numbers in parentheses in the three last columns represent proportions annotated as cancer.

Sample	Pathological annotation	Spots covered by tissue	Genes per spots	Unique transcripts covered by tissue (x 10^4)	Unique transcripts per spot	Area [%]		
						Stroma	*Epithelium*	*Lumen*
1.1	No cancer	432	802	123.4	2857	75.7	11.0	13.3
1.2	Gs 3+3 & PIN3	406	2895	366.2	9019	53.6(0.21)	34.5(0.35)	11.9(0.04)
1.3	Gs 3+3 (outside of area covered by array spots)	629	1377	282.0	4483	82.0(0.82)	8.9(0.19)	9.1(0.05)
2.1	No cancer	547	2869	481.9	8810	70.3	8.3	19.6
2.3	Inflammation	483	346	92.0	1904	81.8	10.4	7.8
2.4	Gs 3+4, suspected Gs	448	790	150.3	3356	79.3(4.0)	9.7(0.15)	11.0(0.01)
3.1	Inflammation	627	1374	152.8	2436	85.6	7.3	7.1
3.2	No cancer	523	553	38.0	727	82.5	8.6	8.9
3.3	Gs 3+4, PIN	501	1200	223.2	4457	83.9(48.16)	11.0(10.22)	5.1(2.54)
4.1	No cancer	688	503	150.9	2194	93.0	2.7	4.3
4.2	Inflammation	324	2062	150.0	4629	77.7	17.5	4.8
4.3	No cancer	302	588	23.4	776	67.9	17.3	14.8

Supplementary Methods

Preparation of quality control arrays and spatially barcoded arrays

In short, for quality control tests poly-T20VN oligonucleotides (IDT) were consistently spread onto Codelink activated microscope glass slides, as per guidelines by the manufacture. Array production for experiments with spatial investigation was described previously[1]. The arrays were designed to have 1007 unique barcoded oligonucleotides with poly-T20VN capture areas. They were printed in areas of size 6200 μm×6600 μm on Codelink activated glass slides with a total of 1007 single spots. In order to keep the orientation, a frame with oligonucleotide (Eurofins) was printed as a border around the barcoded oligonucleotides, containing 148 single spots. Printed spots had a diameter of 100 μm and a center-to-center distance of 200 μm (center to center) from each other.

Poisson factorization core model

Here, we give a high-level description of the core model to perform factor analysis on count matrices such as applicable for spatial gene expression data. The mathematical and computational aspects of the full model, including some extensions, are described in a separate supplement on mathematical methods (**Supplementary Dataset 6**).

We assume that the observed count x_{gs} for gene g in spot s is the sum of count contributions x_{gts} due to T different factors ("cell types"), $x_{gs} = \sum_{t=1}^{T} x_{gts}$, and that these in turn are Poisson distributed, $x_{gts} \sim \mathrm{Pois}(\mu_{gts})$. The Poisson rate parameter μ_{gts} is the product of a gene- and type-dependent gene expression value ϕ_{gt} and a type- and spot-dependent spatial activity value θ_{ts}, $\mu_{gts} = \phi_{gt}\theta_{ts}$. Notably, the gene expression ϕ_{gt} is independent of the spot, while the spatial activity θ_{ts} is independent of the gene. Then, the Poisson factorization core model is given by

$$x_{gs} = \sum_{t} x_{gts} \tag{1}$$

$$x_{gts} \sim \mathrm{Pois}(\mu_{gts}) \tag{2}$$

$$\mu_{gts} = \phi_{gt}\theta_{ts}, \tag{3}$$

where the distributions of the non-negative random variables ϕ_{gt} and θ_{ts} still need specification. A graphical representation of the Poisson factorization core model is displayed in **Supplementary Fig. 29** , and parameters are learned by Monte-Carlo Markov chain (MCMC) sampling.

In Poisson factorization the expected observations are given by the matrix product of the gene expression and spatial activity matrices, $\mathbb{E}[X] = \Phi\Theta$, because $x_{gs} = \sum_t x_{gts} \sim \mathrm{Pois}(\sum_t \mu_{gts}) = \mathrm{Pois}(\sum_t \phi_{gt}\theta_{ts})$ and thus $\mathbb{E}[x_{gs}] = \sum_t \phi_{gt}\theta_{ts}$.

The full model is described in a separate mathematical methods supplement (**Supplementary Dataset 6**) and comprises extensions not mentioned here. The extensions include spot-dependent scaling variables, spatial smoothing, as well as capabilities to perform joint analyses for multiple samples.

Hierarchical clustering for sample 1.2

The bright field image was converted into grayscale in Photoshop and loaded in R using the "jpeg"-library. The image was converted to binary dots using the "base"-library. A virtual grid was generated across the tissue and each binary dot was assigned to a grid-cell using the "sp"- and "raster"-library. The values from the spatial spots were used to generate a linear interpolation, using the "Akima"-library, across the array. Hierarchical clustering was carried out based on Euclidean distance between the spots and each cluster was assigned a color. The colors were further assigned to all grid-cells and each binary dot was assigned the color that corresponded to the grid-cell it was localized in. Areas in between clusters were made transparent using the "scales"-library.

Method description for factor trees

The expression profiles of each factor were used to calculate the Jaccard distance[3] between them. Hierarchical clustering agglomeration method ward.D2 was applied to build the tree (R packages "vegdist", "ape" and "stats")[4–6].

Gene expression analysis for sample 1.2

Intra-replicates were extracted from each region (normal=green color, cancer=red color and PIN=blue color in Supplementary Fig. 4b) within the 1.2 cancer tissue sample, and contained between 4-5 spots. At least 3 sets of intra-replicates were created for each area of interest. Count data was generated with HTSeq-count (version 0.6.1)[13]. The -m parameter was set to union and the count data was imported into the statistical software R. A heatmap was made of the most variable genes, across all regions' replicates, using regularized log (rlog) transformed count values (DESeq2, version 1.12.3)[14]. Principal components were calculated

based on the 500 most variable gene counts after rlog transformation. Individual genes signatures were plotted in concurrence with their location on the spatial array. The sizes of dots are proportional to normalized counts per million (CPM) values for the specific gene for visualization purposes.

Gene expression analysis for sample 3.3

Gene counts were extracted from each spot belonging to the specified conditions (normal, cancer or PIN contained between 3-6 spots per replicate) of interest as determined by the ST analysis. The gene counts were imported into R and CPM values were computed. A matrix was created with rows corresponding to genes and columns corresponding to samples. Row variances were computed across all genes to extract the top 500 genes with highest variance across all regions' replicates. The CPM values for these genes were visualized in a heatmap (Supplementary Fig. 7, Supplementary Table 3).

Immunohistochemistry

Frozen tissue sections stored in -80°C were thawed in RT to be fixated with 3% freshly made paraformaldehyde in TBS for 10 min in RT. Tissues were then permeabilized for 10 minutes in TBS+0.1%Triton-X100, rinsed three times in TBS for 5 minutes/ rinse. Blocking with 2% bovine serum albumin in TBS for 2 hours was performed before the tissues were incubated with primary antibodies overnight at 4°C. After rinsing with 3x5 minutes with TBS the tissues were incubated with the secondary antibodies donkey anti-mouse immunoglobulin G (IgG)-AlexaFluor 568 (1:500 Molecular Probes) and donkey anti-rabbit IgG-AlexaFluor 647 (1:500 Molecular Probes) for 1 hour at RT in darkness. DNA was counterstained with DAPI (Molecular Probes) and slides were mounted with Prolong Gold (Molecular probes). Tissues

were stained with antibodies against SPINK1 (1:50, H00006690-M01, 4D4, Novus), TFF3(1:200, HPA 035464, Sigma), SPON2 (1:100, A-10, st cruz), PGC (1:50, NBP1-91011, Novus), NPY (1:100, ab48789, Abcam), Aquaporin (1:100, ab168387, Abcam), NR4A1 (1:100, ab48789, Abcam), ACPP (1:100, Biologicals), P63(1:150, ab53039, Abcam), Vimentin (1:150, ab8069, Abcam). Fluorescence images were obtained with a Zeiss LSM 780 inverted confocal microscope, using a Plan Apochromat 20×/NA (numerical aperture) 0.7 objective. Tiled images were acquired from optical sections of 5 micrometer.

Pathway analysis in Fig. 3

The gene expression profiles of each factor are the basis for the pathway annotation. We performed outlier detection genewise of the normalised expected values to extract differences in expression between the factors. For each gene in each factor a z-score was calculated based on fitting a normal distribution to the gene's expression in the other factors. We defined significant outliers as genes with a z-score > 2.5 and where the distribution passed the Komolgorov-Smirnov normality test at the 0.05 alpha level. The resulting gene list per factor was submitted to PathwAX[11] on the KEGG database[12] (Fig 3c, Supplementary Table 4)

Comparison of expression in cancer center and cancer periphery for samples 1.2, 2.4, and 3.3

We used the normalized ST-counts per gene and spot. Starting from the size of the area where the "cancer" factor is active, we chose 42 spots (sample 1.2), 403 spots (sample3.3) , and 17 spots (sample 2.4) respectively, comprised of this area.

To ensure quality we removed spots with a log-library size lower than 3 median absolute deviations below the median log-library size (R package "scater")[8]. Additionally, we removed low-abundance genes with zero or near-zero counts. The filtered data set is normalized using the deconvolution method which is based on pool-based size factors and the assumption that most genes are not differentially expressed. The counts from cells were pooled to calculate the size-factors which are used for a cell-specific normalisation (R package "scran")[9]. We utilize the quickCluster method from the R package "scan" to identify optimal pool sizes.

The resulting normalized counts per gene and spot within the tumor region were used to compare expression in the tumor periphery with the center.

Spots located in the periphery and in the center were defined based on the pathologist's annotation and the active cancer factor. Spots with mainly stroma cells were removed. The fold change per gene was calculated as gene expression mean of center spots divided by gene expression mean of periphery spots. P-values per gene were calculated with a two sample t-test[10] at confidence level 0.95. Genes with a p-value < 0.05 were submitted to PathwAX[11] on the KEGG database[12] (Supplementary Tables 8-10).

Pathway analysis in Fig. 5

Ingenuity Pathway Analysis software (Build version 456367M, Content version 39480507 release date 20170914) (Ingenuity Systems, Redwood City, CA) was used to identify significantly enriched pathways. To calculate significance of enrichment (Fisher´s exact test, performed within the software) the reference molecule set was Ingenuity Knowledge Base (Genes only). Input data was extracted from the "Stroma - PTGDS enriched" and "Reactive stroma" factors in Fig. 5, using the top 200 genes (Supplementary Table 5 and 6).

Shiny application

The HE and Cy3 images were aligned and the spots under the tissue stain detected. These spots' coordinates and the corresponding transcript counts were used in the further analysis and image processing. In order to ensure quick data visualization, a Shiny application was built and is freely available at https://spatialtranscriptomics3d.shinyapps.io/STProstateResearch/. The application visualizes spatial gene expression as an interpolation of a regular grid[15] of the coordinate points previously discussed. Then, a tissue mask is placed on top of the interpolated grid to create the final heatmap image presented in the application.

Shiny application access information:

Login with a Google account: Username: 3dstresearch@gmail.com, Password: AIK2017!

Extraction and fragmentation of DNA

DNA was extracted from adjacent sections to each of the twelve sections used for the spatial barcoded array. In order to give a total amount of 100 µm, a total of five sections per bulk sample were cryosectioned at 20 µm. PCa tissue was put in Lysing Matrix D tubes (#116913050, MP Biomedicals) and homogenized in a FastPrep (MP Biomedicals). All the samples were then prepared with the AllPrep DNA/RNA Micro Kit (#80284, Qiagen). DNA extracted from blood was included as germline control using the Gentra Puregene Blood kit (#158445, Qiagen).

Library preparation for whole genome sequencing

Whole genomes libraries were made from extracted DNA (both from tissue and blood) using the NeoPrep Library Prep System (Illumina TruSeq Nano) according to the manufacturer's protocol. Libraries were sequenced with at least 30X (tissue) or 42X (blood) coverage on HiSeqX (HiSeq Control Software 3.3.39/RTA 2.7.1) with a 2x151 setup using HiSeq X SBS chemistry.

DNA sequence alignment

The reads of each sample were aligned with Burrows-Wheeler Aligner (BWA) against the human assembly GRCh38 (ensemble) release 84. BWA *mem* was performed since it is recommended for high-quality queries and longer sequences[16]. The reads in our whole genome sequence (WGS) data have a read length of 151 bp. Furthermore, we used Samtools[17] for converting sam to bam format, sorting, indexing, and for converting the alignments to bed files.

Copy number calling

Copy numbers were inferred with the R package "ReadDepth"[18] based on WGS data of the twelve tissue samples, which was aligned with BWA. Firstly, "ReadDepth" was applied with default values but with annotations computed for our read length of 151 bp and GRCh38. The copy number values were inferred for each sample independently. Secondly, we matched the segments and corresponding copy numbers inferred by "ReadDepth" to GRCh38 release 84 to compare copy number variations (CNVs) in coding regions. We filtered all exons of protein coding genes from the reference genome. If start and end position of an exon is located within a segment, we applied the corresponding copy number. In the rare case of two or more segments that span an exon (1,608 out of 343,705 exons; 0,46%), we assigned two

(1,606 exons) or more (2 exons) copy numbers, however, considered the segment length per copy number within these exons.

For the CNV analysis we needed to know which copy number value is normal. In our samples the peak of base pairs per rounded copy number value for each sample is at 1.8. Therefore, we set a CNV of 1.8 as normal instead of 2.0. The chromosomes X and Y were excluded since here the normal copy number value is one and the borderline for deletions and amplifications is different from the autosomes.

Similarity tree building

The smallest bin size of the calculated segments and copy numbers by "ReadDepth" is 1,200 bp for each of the twelve samples. In other words, each segment start, end, and length of the twelve samples is a multiple of 1,200 bp. Therefore, we sliced the genome in segments of 1,200 bp length and assigned the corresponding copy number values of each sample. This results in approximately 2,5 Mio segments (3,088 Mbp genome length/1,200 bp) and a copy number value vector per segment containing one copy number for each sample. To build the tree we filtered for clear deletions and amplifications by only accepting the CNV vector as input data if the CNV was below 1.6 or above 2.3. Furthermore, we excluded the 1,200 bp segments that contain the centromere since centromeres are difficult to sequence reliably. We also excluded the X and Y chromosomes since their normal CNV value is one. Finally, only unique CNV vectors were chosen to reduce the data space. The data set for the tree contains 287 unique CNV vectors. The tree was built with the R package "Pvclust"[19] based on Euclidean distances, the hierarchical clustering agglomeration method ward.D2, and 1,000 bootstraps.

Copy number variation correlation with gene expression

Copy numbers were inferred as described earlier with he R package "Readdepth" and the segments were mapped to GRCh38 release 84. To display clear deletions and amplification, the copy numbers were corrected by a scaling summand of 0.2 in order to recenter the values (compare Copy number calling). Genes with a break point were excluded.

Additionally, the chromosomes X and Y were excluded since their normal copy number is one. The expression spot mean is presented as relative to the sample expression spot mean and cut at 2000% to emphasize genes with a copy number below or above two. The figures were generated with R package "Scatterplot3d"[20].

Probes and primers

Surface reverse transcription oligonucleotide for quality control experiments
[AmC6]UUUUUGACTCGTAATACGACTCACTATAGGGACACGACGCTCTTCCGATC
TNNNNNNNNTTTTTTTTTTTTTTTTTTTVN

Surface reverse transcription oligonucleotides with spatial barcodes:
[AmC6]UUUUUGACTCGTAATACGACTCACTATAGGGACACGACGCTCTTCCGATC
T[18mer_Spatial_Barcode_1to1007]WSNNWSN

Surface frame oligonucleotide:
[AmC6]AAATTTCGTCTGCTATCGCGCTTCTGTACC

aRNA ligation adapter:
[rApp]AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC[ddC]

Second reverse transcription primer:
GTGACTGGAGTTCAGACGTGTGCTCTTCCGA

PCR primer InPE1.0 primer:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA
TCT

PCR primer InPE2.0 primer:

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

PCR Index primer:

CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTC

Cy3 anti-A probe:

[Cy3]AGATCGGAAGAGCGTCGTGT

Cy3 anti-frame probe:

[Cy3]GGTACAGAAGCGCGATAGCAG

Supplementary References

1. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (80-.).* **353,** 78–82 (2016).
2. van der Maaten, L.J.P. and Hinton, G. E. Visualizing High-Dimensional Data Using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).
3. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. la Société Vaudoise des Sci. Nat.* **37,** 547–579 (1901).
4. Oksanen, J. *et al.* vegan: Community Ecology Package. R package version 2.4-6. (2018).
5. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
6. Team, R. C. R: A language and environment for statistical computing. R Foundation for Statistical Computing.
7. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotech* **33,** 495–502 (2015).
8. McCarthy, D., Wills, Q. & Campbell, K. scater: Single-cell analysis toolkit for gene expression data in R. (2016).
9. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5,** 2122 (2016).
10. WELCH, B. L. The generalisation of student's problems when several different population variances are involved. *Biometrika* **34,** 28–35 (1947).

11. Ogris, C., Helleday, T. & Sonnhammer, E. L. L. PathwAX: a web server for network crosstalk based pathway annotation. *Nucleic Acids Res.* **44,** W105-9 (2016).

12. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28,** 27–30 (2000).

13. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

14. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** (2014).

15. Akima, H. & Gebhardt, A. akima: Interpolation of Irregularly and Regularly Spaced Data. (2015).

16. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589 (2010).

17. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078 (2009).

18. Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads. *PLoS ONE* **6,** (2011).

19. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22,** 1540 (2006).

20. Ligges, U. & Maechler, M. scatterplot3d - An R Package for Visualizing Multivariate Data. *J. Stat. Software; Vol 1, Issue 11* (2003). doi:10.18637/jss.v008.i11

Other Supplementary Files

Supplementary Data 1

Data files and PDFs with the results of the factor analysis for sample 1.2.

Supplementary Data 2

Data files and PDFs with the results of the factor analysis for samples 1.2, 2.4 and 3.3.

Supplementary Data 3

Data files and PDFs with the results of the factor analysis for all samples.

Supplementary Data 4

Data files and PDFs with the results of the factor analysis for samples 3.1 and 4.2.

Supplementary Data 5

Data files and PDFs with the results of the factor analysis for samples 1.3, 2.3, 2.4, and 3.1.

Supplementary Data 6

Supplementary mathematical methods. Full description of the probabilistic method for factor analysis.

Supplementary Data 7

Data files and PDFs with the results of the factor analysis for Patient 1, 2 and 3.

Supplementary Table 2

Results from gene expression analysis for sample 1.2

Supplementary Table 3

Results from gene expression analysis for sample 3.3

Supplementary Table 4

Results from pathway analysis of all ten factors

Supplementary Table 5

Results from Pathway analysis with IPA software (normal stroma)

Supplementary Table 6

Results from Pathway analysis with IPA software (reactive stroma)

Supplementary Table 7

Calculations of stroma, epithelial and lumen of factors in Supplementary Fig. 2

Supplementary Table 8

Results from pathway analysis of center and periphery of cancer in sample 1.2

Supplementary Table 9

Results from pathway analysis of center and periphery of cancer in sample 3.3

Supplementary Table 10

Results from pathway analysis of center and periphery of cancer in sample 2.4